### Title

User-Friendly Covariance Estimation for Heavy-Tailed Distributions




### Journal

STATISTICAL SCIENCE, 34(3)

### ISSN

0883-4237



### Authors

Ke, Yuan

Minsker, Stanislav

Ren, Zhao

et al.



### Publication Date

2019

### DOI

10.1214/19-STS711

Peer reviewed

*Statistical Science*
2019, Vol. 34, No. 3, 454–471
https://doi.org/10.1214/19-STS711


# User-Friendly Covariance Estimation for Heavy-Tailed Distributions

**Yuan Ke, Stanislav Minsker, Zhao Ren, Qiang Sun and Wen-Xin Zhou**

*Abstract.* We provide a survey of recent results on covariance estimation for heavy-tailed distributions. By unifying ideas scattered in the literature, we propose user-friendly methods that facilitate practical implementation. Specifically, we introduce element-wise and spectrum-wise truncation operators, as well as their $M$-estimator counterparts, to robustify the sample covariance matrix. Different from the classical notion of robustness that is characterized by the breakdown property, we focus on the tail robustness which is evidenced by the connection between nonasymptotic deviation and confidence level. The key insight is that estimators should adapt to the sample size, dimensionality and noise level to achieve optimal tradeoff between bias and robustness. Furthermore, to facilitate practical implementation, we propose data-driven procedures that automatically calibrate the tuning parameters. We demonstrate their applications to a series of structured models in high dimensions, including the bandable and low-rank covariance matrices and sparse precision matrices. Numerical studies lend strong support to the proposed methods.

*Key words and phrases:* Covariance estimation, heavy-tailed data, $M$-estimation, nonasymptotics, tail robustness, truncation.

## 1. INTRODUCTION

Covariance estimation serves as a building block for many important statistical learning methods, including principal component analysis, discriminant analysis, clustering analysis and regression analysis, among many others. Recently, estimating structured large covariance matrices, such as bandable, sparse and low-rank matrices, has attracted ever-growing attention in statistics and machine learning (Bickel and Levina 2008a, 2008b, Cai, Ren and Zhou, 2016, Fan, Liao and Liu, 2016). It has broad applications, ranging from functional magnetic resonance imaging (fMRI), analysis of gene expression arrays to risk management and portfolio allocation.

Theoretical properties of large covariance estimators discussed in the literature often hinge heavily on the Gaussian or sub-Gaussian[1] assumption (Vershynin, 2012). See, for example, Theorem 1 of Bickel and Levina (2008a). Such an assumption is typically very restrictive in practice. For example, a recent fMRI study by Eklund, Nichols and Knutsson (2016) reported that most of the common software packages for fMRI analysis, such as SPM and FSL, can result in inflated false-positive rates up to 70% under 5% nominal levels, and questioned a number of fMRI studies among approxi-

*Yuan Ke is Assistant Professor, Department of Statistics, University of Georgia, Athens, Georgia 30602, USA (e-mail: yuan.ke@uga.edu). Stanislav Minsker is Assistant Professor, Department of Mathematics, University of Southern California, Los Angeles, California 90089, USA (e-mail: minsker@usc.edu). Zhao Ren is Assistant Professor, Department of Statistics, University of Pittsburgh, Pittsburgh, Pennsylvania 15260, USA (e-mail: zren@pitt.edu). Qiang Sun is Assistant Professor, Department of Statistical Sciences, University of Toronto, Toronto, Ontario M5S 3G3, Canada (e-mail: qsun@utstat.toronto.edu). Wen-Xin Zhou is Assistant Professor, Department of Mathematics, University of California, San Diego, La Jolla, California 92093, USA (e-mail: wez243@ucsd.edu).*

[1] A random variable $Z$ is said to have sub-Gaussian tails if there exists constants $c_1$ and $c_2$ such that $\mathbb{P}(|Z - \mathbb{E}Z| > t) \leq c_1 \exp(-c_2 t^2)$ for all $t \geq 0$.

mately 40,000 studies according to PubMed. Their results suggested that

> *The principal cause of the invalid cluster inferences is spatial autocorrelation functions that do not follow the assumed Gaussian shape.*

Eklund, Nichols and Knutsson (2016) plotted the empirical versus theoretical spatial autocorrelation functions for several datasets. The empirical autocorrelation functions have much heavier tails compared to their theoretical counterparts under the commonly used assumption of a Gaussian random field, which causes the failure of fMRI inferences. Similar phenomenon has also been discovered in genomic studies (Liu et al., 2003, Purdom and Holmes, 2005) and in quantitative finance (Cont, 2001). It is therefore imperative to develop robust inferential procedures that are less sensitive to the distributional assumptions.

Heavy-tailed distribution is a viable model for data contaminated by outliers that are typically encountered in applications. Due to heavy tailedness, the probability that some observations are sampled far away from the "true" parameter of the population is non-negligible. We refer to these outlying data points as stochastic outliers. A procedure that is robust against such outliers, evidenced by its better finite-sample performance than a non-robust method, is called a *tail-robust* procedure. In this paper, by unifying ideas scattered in the literature, we provide a unified framework for constructing user-friendly tail-robust covariance estimators that admit tight nonasymptotic deviation guarantees under weak moment assumptions. Specifically, we propose element-wise and spectrum-wise truncation operators, as well as their $M$-estimator counterparts, with adaptively chosen robustification parameters. Theoretically, we establish nonasymptotic deviation bounds and demonstrate that the robustification parameters should adapt to the sample size, dimensionality and noise level for optimal tradeoff between bias and robustness. To obtain estimators that are computationally efficient and easily implementable in practice, we propose data-driven schemes to calibrate the tuning parameters, making our proposal user-friendly. Finally, we discuss applications to several structured models in high dimensions, including bandable matrices, low-rank covariance matrices as well as sparse precision matrices. In the supplementary material, we further consider robust covariance estimation and inference under factor models, which might be of independent interest.

Our definition of robustness is different from the conventional perspective under Huber's $\epsilon$-contamination model (Huber, 1964), where the focus has been on developing robust procedures with a high breakdown point. The breakdown point (Hampel, 1971) of an estimator is defined (informally) as the largest proportion of outliers in the data for which the estimator remains stable. Since the seminal work of Tukey (1975), a number of depth-based robust procedures have been developed; see, for example, the papers by Liu (1990), Zuo and Serfling (2000), Mizera (2002) and Salibian-Barrera and Zamar (2002), among others. Another line of work focuses on robust and resistant $M$-estimators, including the least median of squares and least trimmed squares (Rousseeuw, 1984), the S-estimator (Rousseeuw and Yohai, 1984) and the MM-estimator (Yohai, 1987). We refer to Portnoy and He (2000) for a literature review on classical robust statistics, and to Chen, Gao and Ren (2018) for recent developments on nonasymptotic analysis under contamination models.

The rest of the paper is organized as follows. We start with a motivating example in Section 2, which reveals the downsides of the sample covariance matrix. In Section 3, we introduce two types of generic robust covariance estimators and establish their deviation bounds under different norms of interest. The finite-sample performance of the proposed estimators, both element-wise and spectrum-wise, depends on a proper tuning of the robustification parameter that should adapt to the noise level for bias-robustness tradeoff. We also discuss the median-of-means estimator, which is virtually tuning-free at the cost of slightly stronger assumptions. For practical implementation, in Section 4 we propose a data-driven scheme to choose the key tuning parameters. Section 5 presents various applications to estimating structured covariance and precision matrices. Numerical studies are provided in Section 6. We conclude this paper with a discussion in Section 7.

### 1.1 Overview of the Previous Work

In the past several decades, there has been a surge of work on robust covariance estimation in the absence of normality. Examples include the Minimum Covariance Determinant (MCD) estimator, the Minimum Volume Ellipsoid (MVE) estimator, Maronna's (Maronna, 1976) and Tyler's (Tyler, 1987) $M$-estimators of multivariate scatter matrices. We refer to Hubert, Rousseeuw and Van Aelst (2008) for a comprehensive review. Asymptotic properties of these methods have been established for the family of elliptically symmetric distributions; see, for example, Davies (1992), Butler,

Davies and Jhun (1993) and Zhang, Cheng and Singer (2016), among others. However, the aforementioned estimators either rely on parametric assumptions, or impose a shape constraint on the sampling distribution. Under a general setting where neither of these assumptions are made, robust covariance estimation remains a challenging problem.

The work of Catoni (2012) triggered a growing interest in developing tail-robust estimators, which are characterized by tight nonasymptotic deviation analysis, rather than mean squared errors. The current state-of-the-art methods for covariance estimation with heavy-tailed data include those of Catoni (2016), Minsker (2018), Minsker and Wei (2018), Avella-Medina et al. (2018), and Mendelson and Zhivotovskiy (2018). From a spectrum-wise perspective, Catoni (2016) constructed a robust estimator of the Gram and covariance matrices of a random vector $\boldsymbol{X} \in \mathbb{R}^d$ via estimating the quadratic forms $\mathbb{E}\langle \boldsymbol{u}, \boldsymbol{X}\rangle^2$ uniformly over the unit sphere in $\mathbb{R}^d$, and proved error bounds under the operator norm. More recently, Mendelson and Zhivotovskiy (2018) proposed a different robust covariance estimator that admits tight deviation bounds under the finite kurtosis condition. Both constructions, however, involve brute-force search over every direction in a $d$-dimensional $\varepsilon$-net, and thus are computationally intractable. From an element-wise perspective, Avella-Medina et al. (2018) combined robust estimates of the first and second moments to obtain variance estimators. In practice, three potential drawbacks of this approach are: (i) the accumulated error from estimating the first and second moments may cause high variability; (ii) the diagonal variance estimators are not necessarily positive and therefore additional adjustments are required; and (iii) using the cross-validation to calibrate a total number of $O(d^2)$ tuning parameters is computationally expensive.

Building on the ideas of Minsker (2018) and Avella-Medina et al. (2018), we propose user-friendly tail-robust covariance estimators that enjoy desirable finite-sample deviation bounds under weak moment conditions. The constructed estimators only involve simple truncation techniques and are computationally friendly. Through a novel data-driven tuning scheme, we are able to efficiently compute these robust estimators for large-scale problems in practice. These two points distinguish our work from the literature on the topic. The proposed robust procedures serve as building blocks for estimating large structured covariance and precision matrices, and we illustrate their broad applicability in a series of problems.

### 1.2 Notation

We adopt the following notation throughout the paper. For any $0 \le r, s \le \infty$ and a $d \times d$ matrix $\mathbf{A} = (A_{k\ell})_{1\le k,\ell\le d}$, we define the max norm $\|\mathbf{A}\|_{\max} = \max_{1\le k,\ell\le d} |A_{k\ell}|$, the Frobenius norm $\|\mathbf{A}\|_{\mathrm{F}} = (\sum_{1\le k,\ell\le d} A_{k\ell}^2)^{1/2}$ and the operator norm

$$\|\mathbf{A}\|_{r,s} = \sup_{\boldsymbol{u}=(u_1,\ldots,u_d)^\intercal : \|\boldsymbol{u}\|_r = 1} \|\mathbf{A}\boldsymbol{u}\|_s,$$

where $\|\boldsymbol{u}\|_r^r = \sum_{k=1}^d |u_k|^r$ for $r \in (0,\infty)$, $\|\boldsymbol{u}\|_0 = \sum_{k=1}^d I(|u_k| \ne 0)$ and $\|\boldsymbol{u}\|_\infty = \max_{1\le k\le d} |u_k|$. In particular, it holds $\|\mathbf{A}\|_{1,1} = \max_{1\le \ell\le d} \sum_{k=1}^d |A_{k\ell}|$ and $\|\mathbf{A}\|_{\infty,\infty} = \max_{1\le k\le d} \sum_{\ell=1}^d |A_{k\ell}|$. Moreover, we write $\|\mathbf{A}\|_2 := \|\mathbf{A}\|_{2,2}$ for the spectral norm and use $\mathrm{r}(\mathbf{A}) = \mathrm{tr}(\mathbf{A})/\|\mathbf{A}\|_2$ to denote the effective rank of a nonnegative definite matrix $\mathbf{A}$, where $\mathrm{tr}(\mathbf{A}) = \sum_{k=1}^d A_{kk}$ is the trace of $\mathbf{A}$. When $\mathbf{A}$ is symmetric, it is well known that $\|\mathbf{A}\|_2 = \max_{1\le k\le d} |\lambda_k(\mathbf{A})|$ where $\lambda_1(\mathbf{A}) \ge \lambda_2(\mathbf{A}) \ge \cdots \ge \lambda_d(\mathbf{A})$ are the eigenvalues of $\mathbf{A}$. For any matrix $\mathbf{A} \in \mathbb{R}^{d\times d}$ and an index set $J \subseteq \{1,\ldots,d\}^2$, we use $J^{\mathrm{c}}$ to denote the complement of $J$, and $\mathbf{A}_J$ to denote the submatrix of $\mathbf{A}$ with entries indexed by $J$. For a real-valued random variable $X$, let $\mathrm{kurt}(X)$ be the kurtosis of $X$, defined as $\mathrm{kurt}(X) = \mathbb{E}(X-\mu)^4/\sigma^4$, where $\mu = \mathbb{E}X$ and $\sigma^2 = \mathrm{var}(X)$.

## 2. MOTIVATING EXAMPLE: A CHALLENGE OF HEAVY-TAILEDNESS

Suppose that we observe a sample of independent and identically distributed (i.i.d.) copies $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n$ of a random vector $\boldsymbol{X} = (X_1, \ldots, X_d)^\intercal \in \mathbb{R}^d$ with mean $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma} = (\sigma_{k\ell})_{1\le k,\ell\le d}$. To assess the difficulty of mean and covariance estimation for heavy-tailed distributions, we first provide a lower bound for the deviation of the empirical mean under the $\ell_\infty$-norm in $\mathbb{R}^d$.

PROPOSITION 2.1. *For any $\sigma > 0$ and $0 < \delta < (2e)^{-1}$, there exists a distribution in $\mathbb{R}^d$ with mean $\boldsymbol{\mu}$ and covariance matrix $\sigma^2\mathbf{I}_d$ such that the empirical mean $\bar{\boldsymbol{X}} = (1/n)\sum_{i=1}^n \boldsymbol{X}_i$ of i.i.d. observations $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n$ from this distribution satisfies, with probability at least $\delta$,*

$$\|\bar{\boldsymbol{X}} - \boldsymbol{\mu}\|_\infty \ge \sigma\sqrt{\frac{d}{n\delta}}\left(1 - \frac{2e\delta}{n}\right)^{(n-1)/2}. \tag{2.1}$$

The above proposition is a multivariate extension of Proposition 6.2 of Catoni (2012). It provides a lower

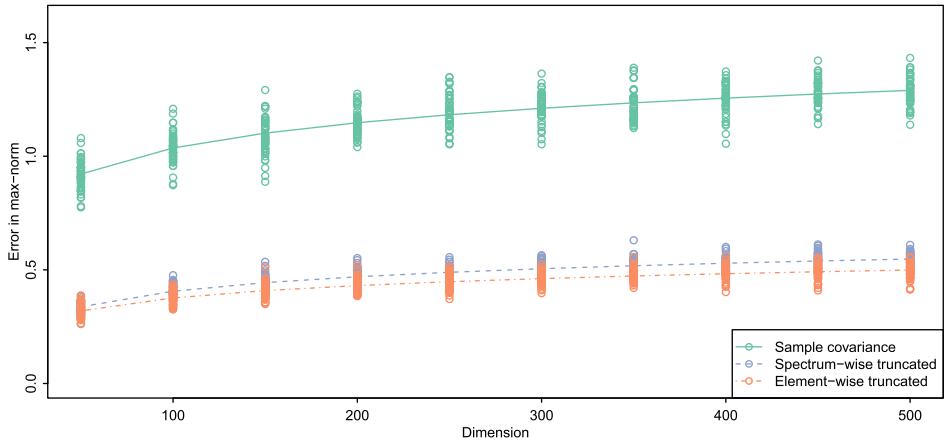


FIG. 1. *Plots of estimation error under max norm versus dimension.*

bound under the $\ell_\infty$-norm for estimating a mean vector via the empirical mean. On the other hand, combining the union bound with Chebyshev's inequality, we obtain that with probability at least $1-\delta$,

$$\|\bar{\boldsymbol{X}} - \boldsymbol{\mu}\|_\infty \le \sigma\sqrt{\frac{d}{n\delta}}.$$

Together, this upper bound and inequality (2.1) show that the worst case deviations of the empirical means grow polynomially in $1/\delta$ under the $\ell_\infty$-norm in the presence of heavy-tailed distributions. As we shall see later, a tail-robust estimator can achieve an exponential-type deviation bound under weak moment conditions.

To demonstrate the practical implications of Proposition 2.1, we perform a toy numerical study on covariance matrix estimation. Let $\boldsymbol{X}_1, \dots, \boldsymbol{X}_n$ be i.i.d. copies of $\boldsymbol{X} = (X_1, \dots, X_d) \in \mathbb{R}^d$, where $X_k$'s are independent and have centered Gamma(3, 1) distribution so that $\boldsymbol{\mu} = \mathbf{0}$ and $\boldsymbol{\Sigma} = 3\mathbf{I}_d$. We compare the performance of three methods: the sample covariance matrix, the element-wise truncated covariance matrix and the spectrum-wise truncated covariance matrix. The latter two are tail-robust covariance estimators that will be introduced in Sections 3.1 and 3.2 respectively. Take $n = 200$ and let $d$ increase from 50 to 500 with a step size of 50. We report the estimation errors under the max norm based on 50 simulations. Figure 1 displays the average (line) and the spread (dots) of estimation errors for each method as the dimension increases. We see that the sample covariance estimator has not only the largest average error but also the highest variability in all the settings. This example demonstrates that the sample covariance matrix suffers from poor finite-sample performance when data are heavy-tailed.

## 3. TAIL-ROBUST COVARIANCE ESTIMATION

### 3.1 Element-Wise Truncated Estimator

We consider the same setting as in the previous section. For mean estimation, the suboptimality of deviations of $\bar{\boldsymbol{X}} = (\bar{X}_1, \dots, \bar{X}_d)^\intercal$ under $\ell_\infty$-norm is due to the fact that the tail probability of $|\bar{X}_k - \mu_k|$ decays only polynomially in the deviation. A simple yet natural idea for improvement is to truncate the data to eliminate outliers introduced by heavy-tailed noises, so that each entry of the resulting estimator exhibits sub-Gaussian tails. To execute this idea, we introduce the following truncation operator, which is closely related to the Huber loss.

DEFINITION 3.1 (Truncation operator). Let $\psi_\tau(\cdot)$ be a truncation operator given by

$$\psi_\tau(u) = \big(|u| \wedge \tau\big)\operatorname{sign}(u), \quad u \in \mathbb{R}, \tag{3.1}$$

where the truncation parameter $\tau > 0$ is also referred to as the *robustification parameter* that trades off bias against robustness.

As an illustration, we assume that $\boldsymbol{\mu} = \mathbf{0}$ whence $\boldsymbol{\Sigma} = \mathbb{E}(\boldsymbol{X}\boldsymbol{X}^\intercal)$. We apply the truncation operator above to each entry of $\boldsymbol{X}_i\boldsymbol{X}_i^\intercal$, and then take the average to obtain

$$\widehat{\sigma}^{\mathcal{T}}_{0,k\ell} = \frac{1}{n}\sum_{i=1}^n \psi_{\tau_{k\ell}}(X_{ik}X_{i\ell}), \quad 1 \le k, \ell \le d,$$

where $\tau_{k\ell} > 0$ are robustification parameters. When the mean vector $\boldsymbol{\mu}$ is unspecified, a straightforward approach is to first estimate the mean vector using existing robust methods (Minsker, 2015, Lugosi and Mendelson, 2019), and then to employ $\widehat{\sigma}^{\mathcal{T}}_{0,k\ell}$ as robust

estimates of the second moments. Estimating the first and second moments separately will unavoidably introduce additional tuning parameters, thus increasing both statistical variability and computational complexity. In what follows, we propose to use the pairwise difference approach to directly estimate variances and covariances, which is free of mean estimation. To the best of our knowledge, the difference-based techniques can be traced back to Rice (1984) and Hall, Kay and Titterington (1990) in the context of bandwidth selection and variance estimation for nonparametric regression.

Let $N := n(n-1)/2$ and define the paired data

$$\begin{aligned}&\{\boldsymbol{Y}_1, \boldsymbol{Y}_2, \ldots, \boldsymbol{Y}_N\}\\ &\quad = \{\boldsymbol{X}_1 - \boldsymbol{X}_2, \boldsymbol{X}_1 - \boldsymbol{X}_3, \ldots, \boldsymbol{X}_{n-1} - \boldsymbol{X}_n\},\end{aligned} \tag{3.2}$$

which are identically distributed from a random vector $\boldsymbol{Y}$ with mean $\mathbf{0}$ and covariance matrix $\operatorname{cov}(\boldsymbol{Y}) = 2\boldsymbol{\Sigma}$. It is easy to check that the sample covariance matrix, $\widehat{\boldsymbol{\Sigma}}^{\mathrm{sam}} = (1/n)\sum_{i=1}^n (\boldsymbol{X}_i - \bar{\boldsymbol{X}})(\boldsymbol{X}_i - \bar{\boldsymbol{X}})^\intercal$ with $\bar{\boldsymbol{X}} = (1/n)\sum_{i=1}^n \boldsymbol{X}_i$, can be expressed as a $U$-statistic

$$\widehat{\boldsymbol{\Sigma}}^{\mathrm{sam}} = \frac{1}{N}\sum_{i=1}^N \boldsymbol{Y}_i \boldsymbol{Y}_i^\intercal/2.$$

Following the argument from the last section, we apply the truncation operator $\psi_\tau$ to $\boldsymbol{Y}_i\boldsymbol{Y}_i^\intercal/2$ entry-wise, and then take the average to obtain

$$\widehat{\sigma}^{\mathcal{T}}_{1,k\ell} = \frac{1}{N}\sum_{i=1}^N \psi_{\tau_{k\ell}}(Y_{ik}Y_{i\ell}/2), \quad 1 \le k, \ell \le d.$$

Concatenating these estimators, we define the element-wise truncated covariance matrix estimator via

$$\widehat{\boldsymbol{\Sigma}}^{\mathcal{T}}_1 = \widehat{\boldsymbol{\Sigma}}^{\mathcal{T}}_1(\boldsymbol{\Gamma}) = \big(\widehat{\sigma}^{\mathcal{T}}_{1,k\ell}\big)_{1\le k,\ell\le d}, \tag{3.3}$$

where $\boldsymbol{\Gamma} = (\tau_{k\ell})_{1\le k,\ell\le d}$ is a symmetric matrix of parameters. $\widehat{\boldsymbol{\Sigma}}^{\mathcal{T}}_1$ can be viewed as a truncated version of the sample covariance matrix $\widehat{\boldsymbol{\Sigma}}^{\mathrm{sam}}$. We assume that $n \ge 2, d \ge 1$ and define $m = \lfloor n/2 \rfloor$, the largest integer not exceeding $n/2$. Moreover, let $\mathbf{V} = (v_{k\ell})_{1\le k,\ell\le d}$ be a symmetric $d \times d$ matrix such that

$$\begin{aligned}v_{k\ell}^2 &= \mathbb{E}(Y_{1k}Y_{1\ell}/2)^2\\ &= \mathbb{E}\big\{(X_{1k} - X_{2k})(X_{1\ell} - X_{2\ell})\big\}^2/4.\end{aligned}$$

THEOREM 3.1. *For any $0 < \delta < 1$, the estimator $\widehat{\boldsymbol{\Sigma}}^{\mathcal{T}}_1 = \widehat{\boldsymbol{\Sigma}}^{\mathcal{T}}_1(\boldsymbol{\Gamma})$ defined in (3.3) with*

$$\boldsymbol{\Gamma} = \sqrt{m/(2\log d + \log\delta^{-1})}\,\mathbf{V} \tag{3.4}$$

*satisfies*

$$\begin{aligned}&\mathbb{P}\Big(\big\|\widehat{\boldsymbol{\Sigma}}^{\mathcal{T}}_1 - \boldsymbol{\Sigma}\big\|_{\max}\\ &\qquad \ge 2\|\mathbf{V}\|_{\max}\sqrt{\frac{2\log d + \log\delta^{-1}}{m}}\Big) \le 2\delta.\end{aligned} \tag{3.5}$$

Theorem 3.1 indicates that, with properly calibrated parameter matrix $\boldsymbol{\Gamma}$, the resulting covariance matrix estimator achieves element-wise tail robustness against heavy-tailed distributions: provided the fourth moments are bounded, each entry of $\widehat{\boldsymbol{\Sigma}}^{\mathcal{T}}_1$ concentrates tightly around its mean so that the maximum error scales as $\sqrt{\log(d)/n} + \sqrt{\log(\delta^{-1})/n}$. Element-wise, we are able to accurately estimate $\boldsymbol{\Sigma}$ at high confidence levels under the constraint that $\log(d)/n$ is small. Implicitly, the dimension $d = d(n)$ is regarded as a function of $n$, and we shall use array asymptotics "$n, d \to \infty$" to characterize large sample behaviors. The finite sample performance, on the other hand, is characterized via nonasymptotic probabilistic bounds with explicit dependence on $n$ and $d$.

REMARK 1. It is worth mentioning that the estimator given in (3.3) and (3.4) is not a genuine sub-Gaussian estimator, in the sense that it depends on the confidence level $1 - \delta$ at which one aims to control the error. More precisely, following the terminology used by Devroye et al. (2016), it is called a $\delta$-dependent sub-Gaussian estimator. Estimators of a similar type include those of Catoni (2012), Minsker (2015), Brownlees, Joly and Lugosi (2015), Hsu and Sabato (2016), Minsker (2018) and Avella-Medina et al. (2018), among others. For univariate mean estimation, Devroye et al. (2016) proposed multiple-$\delta$ mean estimators that satisfy exponential-type concentration bounds uniformly over $\delta \in [\delta_{\min}, 1)$. The idea is to combine a sequence of $\delta$-dependent estimators in a way similar to Lepski's method (Lepskiĭ, 1990).

REMARK 2. Since the element-wise truncated estimator is obtained by treating each covariance $\sigma_{k\ell}$ separately as a univariate parameter, the problem is equivalent to estimation of a large vector given by the concatenation of the columns of $\boldsymbol{\Sigma}$. This type of result is particularly useful for proving upper bounds for sparse covariance and precision estimators in high dimensions; see Section 5. Integrated with $\ell_\infty$-type perturbation bounds, it can also be applied to principle component analysis and factor analysis for heavy-tailed data (Fan et al., 2019). However, when dealing

with large covariance matrices with bandable or low-rank structure, controlling the estimation error under spectral norm is arguably more relevant. A natural idea is then to truncate the spectrum of the sample covariance matrix instead of its entries, which leads to the spectrum-wise truncated estimator defined in the following section.

### 3.2 Spectrum-Wise Truncated Estimator

In this section, we propose and study a covariance estimator that is tail-robust in the spectral norm. To this end, we directly apply the truncation operator to matrices in their spectrum domain. We need the following standard definition of a matrix functional.

DEFINITION 3.2 (Matrix functional). Given a real-valued function $f$ defined on $\mathbb{R}$ and a symmetric $\mathbf{A} \in \mathbb{R}^{K\times K}$ with eigenvalue decomposition $\mathbf{A} = \mathbf{U}\mathbf{\Lambda}\mathbf{U}^\intercal$ such that $\mathbf{\Lambda} = \operatorname{diag}(\lambda_1, \ldots, \lambda_K)$, $f(\mathbf{A})$ is defined as $f(\mathbf{A}) = \mathbf{U} f(\mathbf{\Lambda})\mathbf{U}^\intercal$, where $f(\mathbf{\Lambda}) = \operatorname{diag}(f(\lambda_1), \ldots, f(\lambda_K))$.

Following the same rationale as in the previous section, we propose a spectrum-wise truncated covariance estimator based on the pairwise difference approach:

$$\widehat{\mathbf{\Sigma}}_2^{\mathcal{T}} = \widehat{\mathbf{\Sigma}}_2^{\mathcal{T}}(\tau) = \frac{1}{N}\sum_{i=1}^{N} \psi_\tau(\boldsymbol{Y}_i \boldsymbol{Y}_i^\intercal/2), \tag{3.6}$$

where $\boldsymbol{Y}_i$ are given in (3.3). Note that $\boldsymbol{Y}_i\boldsymbol{Y}_i^\intercal/2$ is a rank-one matrix with eigenvalue $\|\boldsymbol{Y}_i\|_2^2/2$ and the corresponding eigenvector $\boldsymbol{Y}_i/\|\boldsymbol{Y}_i\|_2$. By Definition 3.2, $\widehat{\mathbf{\Sigma}}_2^{\mathcal{T}}$ can be rewritten as

$$\begin{aligned}&\frac{1}{N}\sum_{i=1}^{N} \psi_\tau\left(\frac{1}{2}\|\boldsymbol{Y}_i\|_2^2\right)\frac{\boldsymbol{Y}_i\boldsymbol{Y}_i^\intercal}{\|\boldsymbol{Y}_i\|_2^2}\\ &\quad= \frac{1}{\binom{n}{2}}\sum_{1\le i<j\le n} \psi_\tau\left(\frac{1}{2}\|\boldsymbol{X}_i - \boldsymbol{X}_j\|_2^2\right)\\ &\qquad\times \frac{(\boldsymbol{X}_i - \boldsymbol{X}_j)(\boldsymbol{X}_i - \boldsymbol{X}_j)^\intercal}{\|\boldsymbol{X}_i - \boldsymbol{X}_j\|_2^2}.\end{aligned}$$

This alternative expression renders the computation almost effortless. The following theorem provides an exponential-type concentration inequality for $\widehat{\mathbf{\Sigma}}_2^{\mathcal{T}}$ under operator norm, which is a useful complement to Minsker (2018). Similarly to Theorem 3.1, our next result shows that $\widehat{\mathbf{\Sigma}}_2^{\mathcal{T}}$ achieves exponential-type concentration in the operator norm for heavy-tailed data with finite operator-wise fourth moment, meaning that

$$v^2 = \frac{1}{4}\left\|\mathbb{E}\{(\boldsymbol{X}_1 - \boldsymbol{X}_2)(\boldsymbol{X}_1 - \boldsymbol{X}_2)^\intercal\}^2\right\|_2 \tag{3.7}$$

is finite.

THEOREM 3.2. *For any $0 < \delta < 1$, the estimator $\widehat{\mathbf{\Sigma}}_2^{\mathcal{T}} = \widehat{\mathbf{\Sigma}}_2^{\mathcal{T}}(\tau)$ with*

$$\tau = v\sqrt{\frac{m}{\log(2d) + \log\delta^{-1}}} \tag{3.8}$$

*satisfies, with probability at least $1-\delta$,*

$$\|\widehat{\mathbf{\Sigma}}_2^{\mathcal{T}} - \mathbf{\Sigma}\|_2 \le 2v\sqrt{\frac{\log(2d) + \log\delta^{-1}}{m}}. \tag{3.9}$$

To better recognize this result, note that $v^2$ can be written as

$$\frac{1}{2}\left\|\mathbb{E}\{(\boldsymbol{X} - \boldsymbol{\mu})(\boldsymbol{X} - \boldsymbol{\mu})^\intercal\}^2 + \operatorname{tr}(\mathbf{\Sigma})\mathbf{\Sigma} + 2\mathbf{\Sigma}^2\right\|_2,$$

which is well-defined if the fourth moments $\mathbb{E}(X_k^4)$ are finite. Let

$$K = \sup_{\boldsymbol{u}\in\mathbb{R}^d} \operatorname{kurt}(\boldsymbol{u}^\intercal \boldsymbol{X})$$

be the maximum kurtosis of one-dimensional projections of $\boldsymbol{X}$. Then

$$v^2 \le \|\mathbf{\Sigma}\|_2\{(K+1)\operatorname{tr}(\mathbf{\Sigma})/2 + \|\mathbf{\Sigma}\|_2\}.$$

The following result is a direct consequence of Theorem 3.2: $\widehat{\mathbf{\Sigma}}_2^{\mathcal{T}}$ admits exponential-type concentration for data with finite kurtoses.

COROLLARY 3.1. *Assume that $K = \sup_{\boldsymbol{u}\in\mathbb{R}^d}\operatorname{kurt}(\boldsymbol{u}^\intercal\boldsymbol{X})$ is finite. Then, for any $0 < \delta < 1$, the estimator $\widehat{\mathbf{\Sigma}}_2^{\mathcal{T}} = \widehat{\mathbf{\Sigma}}_2^{\mathcal{T}}(\tau)$ defined in Theorem 3.2 satisfies*

$$\begin{aligned}&\|\widehat{\mathbf{\Sigma}}_2^{\mathcal{T}} - \mathbf{\Sigma}\|_2\\ &\quad\lesssim K^{1/2}\|\mathbf{\Sigma}\|_2\sqrt{\frac{\mathrm{r}(\mathbf{\Sigma})(\log d + \log\delta^{-1})}{n}}\end{aligned} \tag{3.10}$$

*with probability at least $1-\delta$. Here and below, "$\lesssim$" stands for "$\le$" up to an absolute constant.*

REMARK 3. An estimator proposed by Mendelson and Zhivotovskiy (2018) achieves a sharper deviation bound, namely, with $\|\mathbf{\Sigma}\|_2\sqrt{\mathrm{r}(\mathbf{\Sigma})(\log d + \log\delta^{-1})}$ in (3.10) improved to $\|\mathbf{\Sigma}\|_2\sqrt{\mathrm{r}(\mathbf{\Sigma})\log\mathrm{r}(\mathbf{\Sigma})} + \|\mathbf{\Sigma}\|_2\sqrt{\log\delta^{-1}}$; in particular, the second term in the deviation bound is controlled by the spectral norm $\|\mathbf{\Sigma}\|_2$ instead of the possibly much larger $\operatorname{tr}(\mathbf{\Sigma})$. Estimators admitting such recovery guarantees are often called "sub-Gaussian" as they achieve performance similar to the sample covariance obtained from data with multivariate normal distributions. Unfortunately,

the aforementioned estimator is computationally intractable. The question of computational tractability was subsequently resolved by Hopkins (2018) and Cherapanamjeri, Flammarion and Bartlett (2019). The former showed that a polynomial-time algorithm achieves statistically optimal rate under the $\ell_2$-norm, and the latter proposed an estimator that has a significantly faster runtime; in particular, these results apply to covariance estimation with theoretical guarantees under Frobenius norm. Yet it remains an open problem to design a polynomial-time algorithm capable of efficiently computing the estimator proposed by Mendelson and Zhivotovskiy (2018) that achieves near-optimal deviation under the spectral norm.

### 3.3 An $M$-Estimation Viewpoint

In this section, we discuss alternative tail-robust covariance estimators from an $M$-estimation perspective, and study both the element-wise and spectrum-wise truncated estimators. The connection with truncated covariance estimators is discussed at the end of this section. To proceed, we revisit the definition of Huber loss.

DEFINITION 3.3 (Huber loss). The Huber loss $\ell_\tau(\cdot)$ (Huber, 1964) is defined as

$$\ell_\tau(u) = \begin{cases} u^2/2 & \text{if } |u| \le \tau, \\ \tau|u| - \tau^2/2 & \text{if } |u| > \tau, \end{cases} \tag{3.11}$$

where $\tau > 0$ is a robustification parameter similar to that in Definition 3.1.

Compared with the squared error loss, large values of $u$ are down-weighted in the Huber loss, yielding robustness. Generally speaking, minimizing Huber's loss produces a biased estimator of the mean, and parameter $\tau$ can be chosen to control the bias. In other words, $\tau$ quantifies the tradeoff between bias and robustness. As observed by Sun, Zhou and Fan (2019), in order to achieve an optimal tradeoff, $\tau$ should adapt to the sample size, dimension and the noise level.

Starting with the element-wise method, we define the entry-wise estimators

$$\widehat{\sigma}^{\mathcal{H}}_{1,k\ell} = \operatorname*{argmin}_{\theta\in\mathbb{R}} \sum_{i=1}^{N} \ell_{\tau_{k\ell}}(Y_{ik}Y_{i\ell}/2 - \theta), \quad 1 \le k, \ell \le d, \tag{3.12}$$

where $\tau_{k\ell}$ are robustification parameters satisfying $\tau_{k\ell} = \tau_{\ell k}$. When $k = \ell$, even though the minimization is over $\mathbb{R}$, it turns out the solution $\widehat{\sigma}^{\mathcal{H}}_{1,kk}$ is still positive almost surely and therefore provides a reasonable estimator of $\sigma^{\mathcal{H}}_{1,kk}$. To see this, for each $1 \le k \le d$, define $\theta_{0k} = \min_{1\le i\le N} Y_{ik}^2/2$ and note that for any $\tau > 0$ and $\theta \le \theta_{0k}$,

$$\sum_{i=1}^{N} \ell_\tau(Y_{ik}^2/2 - \theta) \ge \sum_{i=1}^{N} \ell_\tau(Y_{ik}^2/2 - \theta_{0k}).$$

It implies that $\widehat{\sigma}^{\mathcal{H}}_{1,kk} \ge \theta_{0k}$, which is strictly positive as long as there are no tied observations. Again, concatenating these marginal estimators, we obtain a Huber-type $M$-estimator

$$\widehat{\mathbf{\Sigma}}^{\mathcal{H}}_1 = \widehat{\mathbf{\Sigma}}^{\mathcal{H}}_1(\mathbf{\Gamma}) = \big(\widehat{\sigma}^{\mathcal{H}}_{1,k\ell}\big)_{1\le k,\ell\le d}, \tag{3.13}$$

where $\mathbf{\Gamma} = (\tau_{k\ell})_{1\le k,\ell\le d}$. The following main result of this section indicates that $\widehat{\mathbf{\Sigma}}^{\mathcal{H}}_1$ achieves tight concentration under the max norm for data with finite fourth moments.

THEOREM 3.3. *Let* $\mathbf{V} = (v_{k\ell})_{1\le k,\ell\le d}$ *be a symmetric matrix with entries*

$$v_{k\ell}^2 = \operatorname{var}\big((X_{1k} - X_{2k})(X_{1\ell} - X_{2\ell})/2\big). \tag{3.14}$$

*For any* $0 < \delta < 1$, *the covariance estimator* $\widehat{\mathbf{\Sigma}}^{\mathcal{H}}_1$ *given in* (3.13) *with*

$$\mathbf{\Gamma} = \sqrt{\frac{m}{2\log d + \log\delta^{-1}}}\mathbf{V} \tag{3.15}$$

*satisfies*

$$\mathbb{P}\bigg(\big\|\widehat{\mathbf{\Sigma}}^{\mathcal{H}}_1 - \mathbf{\Sigma}\big\|_{\max} \ge 4\|\mathbf{V}\|_{\max}\sqrt{\frac{2\log d + \log\delta^{-1}}{m}}\bigg) \le 2\delta \tag{3.16}$$

*as long as* $m \ge 8\log(d^2\delta^{-1})$.

The $M$-estimator counterpart of the spectrum-truncated covariance estimator was first proposed by Minsker (2018) using a different robust loss function, and extended by Minsker and Wei (2018) to more general framework of $U$-statistics. In line with the previous element-wise $M$-estimator, we restrict our attention to the Huber loss and consider

$$\widehat{\mathbf{\Sigma}}^{\mathcal{H}}_2 \in \operatorname*{argmin}_{\mathbf{M}\in\mathbb{R}^{d\times d}:\mathbf{M}=\mathbf{M}^\intercal} \operatorname{tr}\bigg\{\frac{1}{N}\sum_{i=1}^{N}\ell_\tau(\boldsymbol{Y}_i\boldsymbol{Y}_i^\intercal/2 - \mathbf{M})\bigg\}, \tag{3.17}$$

which is a natural robust variant of the sample covariance matrix

$$\widehat{\mathbf{\Sigma}}^{\text{sam}} = \operatorname*{argmin}_{\mathbf{M}\in\mathbb{R}^{d\times d}:\mathbf{M}=\mathbf{M}^\intercal} \operatorname{tr}\bigg\{\frac{1}{N}\sum_{i=1}^{N}(\boldsymbol{Y}_i\boldsymbol{Y}_i^\intercal/2 - \mathbf{M})^2\bigg\}.$$

Define the $d \times d$ matrix $\mathbf{S}_0 = \mathbb{E}\{(\boldsymbol{X}_1 - \boldsymbol{X}_2)(\boldsymbol{X}_1 - \boldsymbol{X}_2)^\top/2 - \boldsymbol{\Sigma}\}^2$ that satisfies

$$\mathbf{S}_0 = \frac{\mathbb{E}\{(\boldsymbol{X} - \boldsymbol{\mu})(\boldsymbol{X} - \boldsymbol{\mu})^\top\}^2 + \operatorname{tr}(\boldsymbol{\Sigma})\boldsymbol{\Sigma}}{2}.$$

The following result is modified from Corollary 4.1 of Minsker and Wei (2018).

THEOREM 3.4. *Assume that there exists some* $K > 0$ *such that* $\sup_{\boldsymbol{u} \in \mathbb{R}^d} \operatorname{kurt}(\boldsymbol{u}^\top \boldsymbol{X}) \le K$. *Then for any* $0 < \delta < 1$ *and* $v \ge \|\mathbf{S}_0\|_2^{1/2}$, *the M-estimator* $\widehat{\boldsymbol{\Sigma}}_2^{\mathcal{H}}$ *with* $\tau = v\sqrt{m/(2\log d + 2\log \delta^{-1})}$ *satisfies*

$$\|\widehat{\boldsymbol{\Sigma}}_2^{\mathcal{H}} - \boldsymbol{\Sigma}\|_2 \le C_1 v \sqrt{\frac{\log d + \log \delta^{-1}}{m}} \tag{3.18}$$

*with probability at least* $1 - 5\delta$ *as long as* $n \ge C_2 K \cdot \mathrm{r}(\boldsymbol{\Sigma})(\log d + \log \delta^{-1})$, *where* $C_1, C_2 > 0$ *are absolute constants*.

To solve the convex optimization problem (3.17), Minsker and Wei (2018) proposed the following gradient descent algorithm: starting with an initial estimator $\widehat{\boldsymbol{\Sigma}}^{(0)}$, at iteration $t = 1, 2, \ldots,$ compute

$$\widehat{\boldsymbol{\Sigma}}^{(t)} = \widehat{\boldsymbol{\Sigma}}^{(t-1)} - \frac{1}{N}\sum_{i=1}^{N} \psi_\tau(\boldsymbol{Y}_i \boldsymbol{Y}_i^\top/2 - \widehat{\boldsymbol{\Sigma}}^{(t-1)}),$$

where $\psi_\tau$ is given in (3.1). From this point of view, the truncated estimator $\widehat{\boldsymbol{\Sigma}}_2^{\mathcal{T}}$ given in (3.6) can be viewed as the first step of the gradient descent iteration for solving optimization problem (3.17) initiated at $\widehat{\boldsymbol{\Sigma}}^{(0)} = \mathbf{0}$. This procedure enjoys a nice contraction property, as demonstrated by Lemma 3.2 of Minsker and Wei (2018). However, since the difference matrix $\boldsymbol{Y}_i \boldsymbol{Y}_i^\top/2 - \widehat{\boldsymbol{\Sigma}}^{(t-1)}$ for each $t$ is no longer rank-one, we need to perform a singular value decomposition to compute the matrix $\psi_\tau(\boldsymbol{Y}_i \boldsymbol{Y}_i^\top/2 - \widehat{\boldsymbol{\Sigma}}^{(t-1)})$ for $i = 1, \ldots, N$.

We end this section with a discussion of the similarities and differences between $M$-estimators and estimators defined via truncation. Both types of estimators achieve tail robustness through a bias-robustness tradeoff, either element-wise or spectrum-wise. However (informally speaking), $M$-estimators truncate symmetrically around the true expectation as shown in (3.12) and (3.17), while the truncation-based estimators truncate around zero as in (3.3) and (3.6). Due to smaller bias, $M$-estimators are expected to outperform the simple truncation estimators. However, since the optimal choice of the robustification parameter is often much larger than the population moments in magnitude, either element-wise or spectrum-wise, the difference between truncation estimators and $M$-estimators becomes insignificant when the sample size $n$ is large. Therefore, we advocate using the simple truncated estimator primarily due to its simplicity and computational efficiency.

### 3.4 Median-of-Means Estimator

Truncation-based approaches described in the previous sections require knowledge of robustification parameters $\tau_{k\ell}$. Adaptation and tuning of these parameters will be discussed in Section 4 below. Here, we suggest another method that does not need any tuning but requires stronger assumptions, namely, existence of moments of order six. This method is based on the median-of-means (MOM) technique (Nemirovsky and Yudin, 1983, Devroye et al., 2016, Minsker and Strawn, 2017). To this end, assume that the index set $\{1, \ldots, n\}$ is partitioned into $k$ disjoint groups $G_1, \ldots, G_k$ (partitioning scheme is assumed to be independent of $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n$) such that the cardinalities $|G_j|$ satisfy $||G_j| - \frac{n}{k}| \le 1$ for $j = 1, \ldots, k$. For each $j = 1, \ldots, k$, let $\bar{\boldsymbol{X}}_{G_j} = (1/|G_j|)\sum_{i \in G_j} \boldsymbol{X}_i$ and

$$\widehat{\boldsymbol{\Sigma}}^{(j)} = \frac{1}{|G_j|}\sum_{i \in G_j} (\boldsymbol{X}_i - \bar{\boldsymbol{X}}_{G_j})(\boldsymbol{X}_i - \bar{\boldsymbol{X}}_{G_j})^\top$$

be the sample covariance evaluated over the data in group $j$. Then, for all $1 \le \ell, m \le d$, the MOM estimator of $\sigma_{\ell m}$ is defined via

$$\widehat{\sigma}_{\ell m}^{\mathrm{MOM}} = \operatorname{median}\{\widehat{\sigma}_{\ell m}^{(1)}, \ldots, \widehat{\sigma}_{\ell m}^{(k)}\},$$

where $\widehat{\sigma}_{\ell m}^{(j)}$ is the entry in the $\ell$th row and $m$th column of $\widehat{\boldsymbol{\Sigma}}^{(j)}$. This leads to

$$\widehat{\boldsymbol{\Sigma}}^{\mathrm{MOM}} = (\widehat{\sigma}_{\ell m}^{\mathrm{MOM}})_{1 \le \ell, m \le d}.$$

Let $\Delta_{\ell m}^2 = \operatorname{Var}((X_\ell - \mathbb{E}X_\ell)(X_m - \mathbb{E}X_m))$ for $1 \le \ell, m \le d$. The following result provides a deviation bound for the MOM estimator $\widehat{\boldsymbol{\Sigma}}^{\mathrm{MOM}}$ under the max norm.

THEOREM 3.5. *Assume that* $\min_{\ell,m} \Delta_{\ell m}^2 \ge c_\ell > 0$ *and* $\max_{1 \le k \le d} \mathbb{E}|X_k - \mathbb{E}X_k|^6 \le c_u < \infty$. *Then, there exists* $C_0 > 0$ *depending only on* $(c_\ell, c_u)$ *such that*

$$\mathbb{P}\Bigg( \|\widehat{\boldsymbol{\Sigma}}^{\mathrm{MOM}} - \boldsymbol{\Sigma}\|_{\max} \ge 3\max_{\ell,m} \Delta_{\ell m} \Bigg\{ \sqrt{\frac{\log(d+1) + \log \delta^{-1}}{n}} + C_0 \frac{k}{n} \Bigg\} \Bigg) \le 2\delta \tag{3.19}$$

*for all* $\delta$ *satisfying* $\sqrt{\{\log(d+1)+\log\delta^{-1}\}/k} + C_0\sqrt{k/n} \le 0.33$.

REMARK 4.

1. The only user-defined parameter in the definition of $\widehat{\mathbf{\Sigma}}^{\mathrm{MOM}}$ is the number of subgroups $k$. The bound above shows that, provided $k \ll \sqrt{n}$ (say, one could set $k = \sqrt{n}/\log n$), the term $C_0 k/n$ in (3.19) is of smaller order, and we obtain an estimator that admits tight deviation bounds for a wide range of $\delta$. In this sense, estimator $\widehat{\mathbf{\Sigma}}^{\mathrm{MOM}}$ is essentially a multiple-$\delta$ estimator (Devroye et al., 2016); see Remark 1.

2. Application of the MOM construction to large covariance estimation problems has been explored by Avella-Medina et al. (2018). However, the results obtained therein are insufficient to conclude that MOM estimators are truly "tuning-free". Under a bounded fourth moment assumption, Avella-Medina et al. (2018) derived a deviation bound (under max norm) for the element-wise median-of-means estimator with the number of partitions depending on a prespecified confidence level parameter.

## 4. AUTOMATIC TUNING OF ROBUSTIFICATION PARAMETERS

For all the proposed tail-robust estimators besides the median-of-means, the robustification parameter needs to adapt to the sample size, dimensionality and noise level in order to achieve optimal tradeoff between bias and robustness in finite samples. An intuitive idea is to use cross-validation or the Lepski's method (Lepski and Spokoiny, 1997, Minsker, 2018). However both approaches are computationally expensive. In this section, we propose tuning-free approaches for constructing both truncated and $M$-estimators that have low computational costs. Our nonasymptotic analysis provides useful guidance on the choice of key tuning parameters.

### 4.1 Adaptive Truncated Estimator

We first introduce a data-driven procedure that automatically tunes the robustification parameters in the element-wise truncated covariance estimator. This procedure is motivated by the theoretical properties established in Theorem 3.1. To avoid notational clutter, we fix $1 \le k \le \ell \le d$ and define $\{Z_1, \dots, Z_N\} = \{Y_{1k}Y_{1\ell}/2, \dots, Y_{Nk}Y_{N\ell}/2\}$ such that $\sigma_{k\ell} = \mathbb{E}Z_1$. Then $\widehat{\sigma}^{\mathcal{T}}_{1,k\ell}$ can be written as $(1/N)\sum_{i=1}^N \psi_{\tau_{k\ell}}(Z_i)$. In view of (3.4), an "ideal" choice of $\tau_{k\ell}$ is

$$\tau_{k\ell} = v_{k\ell}\sqrt{\frac{m}{2\log d + t}} \quad \text{with } v_{k\ell}^2 = \mathbb{E}Z_1^2, \tag{4.1}$$

where $t = \log\delta^{-1} \ge 1$ is prespecified to control the confidence level and will be discussed later. A naive estimator of $v_{k\ell}^2$ is the empirical second moment $(1/N)\sum_{i=1}^N Z_i^2$, which tends to overestimate the true value when data have high kurtoses. Intuitively, a well-chosen $\tau_{k\ell}$ makes $(1/N)\sum_{i=1}^N \psi_{\tau_{k\ell}}(Z_i)$ a good estimator of $\mathbb{E}Z_1$, and meanwhile, we expect the empirical truncated second moment $(1/N)\sum_{i=1}^N \psi_{\tau_{k\ell}}^2(Z_i) = (1/N)\sum_{i=1}^N (Z_i^2 \wedge \tau_{k\ell}^2)$ to be a reasonable estimate of $\mathbb{E}Z_1^2$. Plugging this empirical truncated second moment into (4.1) yields

$$\frac{1}{N}\sum_{i=1}^N \frac{(Z_i^2 \wedge \tau^2)}{\tau^2} = \frac{2\log d + t}{m}, \quad \tau > 0. \tag{4.2}$$

We then solve the above equation to obtain $\widehat{\tau}_{k\ell}$, a data-driven choice of $\tau_{k\ell}$. By Proposition 3 in Wang et al. (2018), equation (4.2) has a unique solution as long as $2\log d + t < (m/N)\sum_{i=1}^N I\{Z_i \ne 0\}$. We characterize the theoretical properties of this tuning method in a companion paper (Chen and Zhou, 2019).

Regarding the choice of $t = \log\delta^{-1}$: on the one hand, because it controls the confidence level according to (3.5), we shall let $t = t_n$ be sufficiently large so that the estimator concentrates around the true value with high probability. On the other hand, $t$ also appears in the deviation bound that corresponds to the width of the confidence interval, so it should not grow too fast as a function of $n$. In practice, we recommend using $t = \log n$ (or equivalently, $\delta = n^{-1}$), a slowly varying function of $n$.

To implement the spectrum-wise truncated covariance estimator in practice, note that there is only one tuning parameter whose theoretically optimal scale is

$$\frac{1}{2}\big\|\mathbb{E}\{(\boldsymbol{X}_1 - \boldsymbol{X}_2)(\boldsymbol{X}_1 - \boldsymbol{X}_2)^{\mathsf{T}}\}^2\big\|_2^{1/2}\sqrt{\frac{m}{\log(2d)+t}}.$$

Motivated by the data-driven tuning scheme for the element-wise estimator, we choose $\tau$ by solving the equation

$$\left\|\frac{1}{\tau^2 N}\sum_{i=1}^N\Big(\frac{\|\boldsymbol{Y}_i\|_2^2}{2} \wedge \tau\Big)^2 \frac{\boldsymbol{Y}_i\boldsymbol{Y}_i^{\mathsf{T}}}{\|\boldsymbol{Y}_i\|_2^2}\right\|_2 = \frac{\log(2d)+t}{m},$$

where as before we take $t = \log n$.

### 4.2 Adaptive Huber-Type $M$-Estimator

To construct a data-driven approach that automatically tunes the adaptive Huber estimator, we follow the same rationale from the previous subsection. Since the optimal $\tau_{k\ell}$ now depends on $\mathrm{var}(Z_1)$ instead of the second moment $\mathbb{E}Z_1^2$, it is therefore conservative to directly apply the above data-driven method in this case. Instead, we propose to estimate $\tau_{k\ell}$ and $\sigma_{k\ell}$ simultaneously by solving the following system of equations

$$f_1(\theta,\tau)=\frac{1}{N}\sum_{i=1}^{N}\frac{\{(Z_i-\theta)^2\wedge\tau^2\}}{\tau^2}-\frac{2\log d+t}{n}=0, \tag{4.3a}$$

$$f_2(\theta,\tau)=\sum_{i=1}^{N}\psi_\tau(Z_i-\theta)=0, \tag{4.3b}$$

for $\theta\in\mathbb{R}$ and $\tau>0$. Via a similar argument, it can be shown that the equation $f_1(\theta,\cdot)=0$ has a unique solution as long as $2\log d+t<(n/N)\sum_{i=1}^{N}I\{Z_i\neq\theta\}$; for any $\tau>0$, the equation $f_2(\cdot,\tau)=0$ also has a unique solution. Starting with an initial estimate $\theta^{(0)}=(1/N)\sum_{i=1}^{N}Z_i$, which is the sample variance estimator of $\sigma_{k\ell}$, we iteratively solve $f_1(\theta^{(s-1)},\tau^{(s)})=0$ and $f_2(\theta^{(s)},\tau^{(s)})=0$ for $s=1,2,\ldots$ until convergence. The resultant estimator, denoted by $\widehat{\sigma}_{3,k\ell}^{\mathcal{H}}$ with slight abuse of notation, is then referred to as the adaptive Huber estimator of $\sigma_{k\ell}$. We then obtain the data-adaptive Huber covariance matrix estimator as $\widehat{\mathbf{\Sigma}}_3^{\mathcal{H}}=(\widehat{\sigma}_{3,k\ell}^{\mathcal{H}})_{1\le k,\ell\le d}$. Algorithm 1 presents the summary of this data-driven approach.

**Algorithm 1** Data-adaptive huber covariance matrix estimation

**Input** Data vectors $\boldsymbol{X}_i\in\mathbb{R}^d$ $(i=1,\ldots,n)$, tolerance level $\epsilon$ and maximum iteration $S_{\max}$.

**Output** Data-adaptive Huber covariance matrix estimator $\widehat{\mathbf{\Sigma}}_3^{\mathcal{H}}=(\widehat{\sigma}_{3,k\ell}^{\mathcal{H}})_{1\le k,\ell\le d}$.

Compute pairwise differences $\boldsymbol{Y}_1=\boldsymbol{X}_1-\boldsymbol{X}_2,\boldsymbol{Y}_2=\boldsymbol{X}_1-\boldsymbol{X}_3,\ldots,\boldsymbol{Y}_N=\boldsymbol{X}_{n-1}-\boldsymbol{X}_n$, where $N=n(n-1)/2$.
  **for** $1\le k\le\ell\le d$ **do**
    $\theta^{(0)}=(2N)^{-1}\sum_{i=1}^{N}Y_{ik}Y_{i\ell}$.
    **for** $s=1,\ldots,S_{\max}$ **do**
      $\tau^{(s)}\leftarrow$ solution of $f_1(\theta^{(s-1)},\cdot)=0$.
      $\theta^{(s)}\leftarrow$ solution of $f_2(\cdot,\tau^{(s)})=0$.
      **if** $|\theta^{(s)}-\theta^{(s-1)}|<\epsilon$ **break**
    **stop** $\widehat{\sigma}_{3,\ell k}^{\mathcal{H}}=\widehat{\sigma}_{3,k\ell}^{\mathcal{H}}=\theta^{(S_{\max})}$.
  **stop**
**return** $\widehat{\mathbf{\Sigma}}_3^{\mathcal{H}}=(\widehat{\sigma}_{3,k\ell}^{\mathcal{H}})_{1\le k,\ell\le d}$.

## 5. APPLICATIONS TO STRUCTURED MATRIX ESTIMATION

The robustness properties of the element-wise and spectrum-wise truncation estimators are demonstrated in Theorems 3.1 and 3.2. In particular, the exponential-type concentration bounds are essential for establishing reasonable estimators for high-dimensional structured covariance and precision matrices. In this section, we apply the proposed generic robust methods to the estimation of bandable and low-rank covariance matrices as well as sparse precision matrices.

### 5.1 Bandable Covariance Matrix Estimation

Motivated by applications to climate studies and spectroscopy in which the index set of variables $\boldsymbol{X}=(X_1,\ldots,X_d)^{\mathsf{T}}$ admits a natural order, one can expect that a large "distance" $|k-\ell|$ implies near-independence. We characterize this feature by the following class of bandable covariance matrices considered by Bickel and Levina (2008a) and by Cai, Zhang and Zhou (2010):

$$\mathcal{F}_\alpha(M_0,M)=\Big\{\mathbf{\Sigma}=(\sigma_{k\ell})_{1\le k,\ell\le d}\in\mathbb{R}^{d\times d}:\lambda_1(\mathbf{\Sigma})\le M_0,\ \max_{1\le\ell\le d}\sum_{k:|k-\ell|>m}|\sigma_{k\ell}|\le\frac{M}{m^\alpha}\text{ for all }m\Big\}. \tag{5.1}$$

Here $M_0,M$ are regarded as universal constants and the parameter $\alpha$ specifies the decay rate of $\sigma_{k\ell}$ to zero as $\ell\to\infty$ for each row.

When $\boldsymbol{X}$ follows sub-Gaussian distribution, Cai, Zhang and Zhou (2010) proposed a minimax-optimal estimator over $\mathcal{F}_\alpha(M_0,M)$ under the spectral norm. Specifically, they proposed a tapering estimator $\widehat{\mathbf{\Sigma}}_m^{\mathrm{tap}}=(\widehat{\sigma}_{k\ell}\cdot\omega_{|k-\ell|})$, where the positive integer $m\le d$ specifies the bandwidth, $\omega_q=1$, $2-2q/m$, $0$, when $q\le m/2$, $m/2<q\le m$, $q>m$, respectively. $\widehat{\mathbf{\Sigma}}^{\mathrm{sam}}=(\widehat{\sigma}_{k\ell})_{1\le k,\ell\le d}$ denotes the sample covariance. With the optimal choice of bandwidth $m\asymp\min\{n^{1/(2\alpha+1)},d\}$, Cai, Zhang and Zhou (2010) showed that $\widehat{\mathbf{\Sigma}}_m^{\mathrm{tap}}$ achieves the minimax rate of convergence $\{\sqrt{\log(d)/n}+n^{-\alpha/(2\alpha+1)}\}\wedge\sqrt{d/n}$ under the spectral norm.

To obtain a root-$n$ consistent covariance estimator, we expect the coordinates of $\boldsymbol{X}$ to have at least fi-

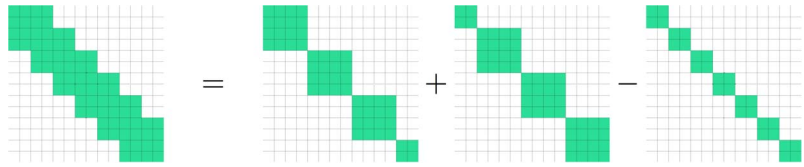

FIG. 2. *Motivation of our estimator of bandable covariance matrices.*

nite fourth moments. Under this condition, it is unclear whether the optimal rate can be achieved over $\mathcal{F}_\alpha(M_0, M)$ without imposing additional distributional assumptions, such as the elliptical symmetry (Mitra and Zhang, 2014, Chen, Gao and Ren, 2018). Estimators that naively use the sample covariance will inherit its sensitivity to outliers. Recall the definition of $\widehat{\mathbf{\Sigma}}_2^{\mathcal{T}}$ in (3.6); a simple idea is to replace the sample covariance by a spectrum-wise truncated estimator $\widehat{\mathbf{\Sigma}}_2^{\mathcal{T}}$ in the first step, to which the tapering procedure can be applied. However, such an estimator is not optimal: indeed, the analysis of a tapering estimator requires each small principal submatrix of the initial estimator to be highly concentrated around the population object. Suppose that we truncate the $\ell_2$-norm of the entire vector $\boldsymbol{Y}_i$ at a level $\tau$ scaling with $\operatorname{tr}(\mathbf{\Sigma})$. For each subset $J \subseteq \{1, \dots, d\}$, let $\boldsymbol{Y}_{iJ}$ be the subvector of $\boldsymbol{Y}_i$ indexed by $J$. Then the corresponding principal submatrix

$$\frac{1}{N}\sum_{i=1}^{N} \psi_\tau\Big(\frac{1}{2}\|\boldsymbol{Y}_i\|_2^2\Big)\frac{\boldsymbol{Y}_{iJ}\boldsymbol{Y}_{iJ}^\intercal}{\|\boldsymbol{Y}_i\|_2^2}$$

is not an ideal robust estimator of $\mathbf{\Sigma}_{JJ}$ because the "optimal" $\tau$ in this case should scale with $\operatorname{tr}(\mathbf{\Sigma}_{JJ})$ rather than $\operatorname{tr}(\mathbf{\Sigma})$. This explains why directly applying the tapering procedure to $\widehat{\mathbf{\Sigma}}_2^{\mathcal{T}}$ is not ideal.

In what follows, we propose an optimal robust covariance estimator based on the spectrum-wise truncation technique introduced in Section 3.2. First, we introduce some notation. Let $\boldsymbol{Z}_i^{(p,q)} = (Y_{i,p}, Y_{i,p+1}, \dots, Y_{i,p+q-1})^\intercal$ be a subvector of $\boldsymbol{Y}_i$ given in (3.3). Accordingly, define the truncated estimator of the principal submatrix of $\mathbf{\Sigma}$ as

$$\begin{aligned}\widehat{\mathbf{\Sigma}}_2^{(p,q),\mathcal{T}} &= \widehat{\mathbf{\Sigma}}_2^{(p,q),\mathcal{T}}(\tau)\\ &= \frac{1}{N}\sum_{i=1}^{N} \psi_\tau\big(\boldsymbol{Z}_i^{(p,q)}\boldsymbol{Z}_i^{(p,q)\intercal}/2\big),\end{aligned} \tag{5.2}$$

where $\tau$ is as in (3.8) with $d$ replaced by $q$ and $v = \|\mathbb{E}\{\boldsymbol{Z}_1^{(p,q)}\boldsymbol{Z}_1^{(p,q)\intercal}\}^2\|_2/4$. Moreover, we define an operator that embeds a small matrix into a large zero matrix: for a $q \times q$ matrix $\mathbf{A} = (a_{k\ell})_{1\le k,\ell\le q}$, define the $d \times d$ matrix $\mathbf{E}_p^d(\mathbf{A}) = (b_{k\ell})_{1\le k,\ell\le d}$, where $p$ indicates the location and

$$b_{k\ell} = \begin{cases} a_{k-p+1,\ell-p+1} & \text{if } p \le k, \ell \le p+q-1,\\ 0 & \text{otherwise.}\end{cases}$$

Our final robust covariance estimator is then defined as

$$\begin{aligned}\widehat{\mathbf{\Sigma}}_q = &\sum_{j=-1}^{\lceil (d-1)/q\rceil} \mathbf{E}_{jq+1}^d\big(\widehat{\mathbf{\Sigma}}_2^{(jq+1,2q),\mathcal{T}}\big)\\ &- \sum_{j=0}^{\lceil (d-1)/q\rceil} \mathbf{E}_{jq+1}^d\big(\widehat{\mathbf{\Sigma}}_2^{(jq+1,q),\mathcal{T}}\big).\end{aligned} \tag{5.3}$$

The idea behind the construction above is that a bandable covariance matrix in $\mathcal{F}_\alpha(M_0, M)$ can be approximately decomposed into several principal submatrices of size $2q$ and $q$, as shown in Figure 2. Using spectrum-wise truncated estimators $\widehat{\mathbf{\Sigma}}_2^{(p,q),\mathcal{T}}$ and $\widehat{\mathbf{\Sigma}}_2^{(p,2q),\mathcal{T}}$ to estimate the corresponding principal submatrices in this decomposition leads to the proposed estimator $\widehat{\mathbf{\Sigma}}_q$.

This construction is different from the literature where the banding or tapering procedure is directly applied to an initial estimator, say the sample covariance matrix (Bickel and Levina, 2008a, Cai, Zhang and Zhou, 2010). It is worth mentioning that a similar robust estimator can be constructed following the idea of Cai, Zhang and Zhou (2010), which differs from our proposal. Computationally, our estimator evaluates as many as $O(d/q)$ matrices of size $q \times q$ (or $2q \times 2q$), while the method developed by Cai, Zhang and Zhou (2010) computes as many as $O(d)$ such matrices.

The following result shows that the estimator defined in (5.3) achieves near-optimal rate of convergence under the spectral norm as long as $\boldsymbol{X}$ has uniformly bounded fourth moments. The proof is deferred to the supplementary material.

THEOREM 5.1. *Assume that* $\mathbf{\Sigma} \in \mathcal{F}_\alpha(M_0, M)$ *and* $\sup_{\boldsymbol{u}\in\mathbb{S}^{d-1}} \operatorname{kurt}(\boldsymbol{u}^\intercal \boldsymbol{X}) \le M_1$ *for some constant* $M_1 > 0$. *For any* $c_0 > 0$, *take* $\delta = (n^{c_0} d)^{-1}$ *in the definition of* $\tau$ *for constructing principal submatrix estimators* $\widehat{\mathbf{\Sigma}}_2^{(p,q),\mathcal{T}}$ *in* (5.2). *Then, with a bandwidth* $q \asymp$

$\{n/\log(nd)\}^{1/(2\alpha+1)} \wedge d$, *the estimator* $\widehat{\mathbf{\Sigma}}_q$ *defined in* (5.3) *is such that with probability at least* $1 - 2n^{-c_0}$,

$$\|\widehat{\mathbf{\Sigma}}_q - \mathbf{\Sigma}\|_2 \le C \min\left\{\left(\frac{\log(nd)}{n}\right)^{\alpha/(2\alpha+1)}, \sqrt{\frac{d \cdot \log(nd)}{n}}\right\},$$

*where* $C > 0$ *is a constant depending only on* $M, M_0, M_1, c_0$.

According to the minimax lower bounds established by Cai, Zhang and Zhou (2010), up to a logarithmic term our robust estimator achieves the optimal rate of convergence that is enjoyed by the tapering estimator when the data are sub-Gaussian. Our estimator is not fully data-driven, because the optimal choice of the bandwidth $q$ depends on the unknown parameter $\alpha$. We refer to Liu and Ren (2018) for a Lepski-type adaptive procedure.

### 5.2 Low-Rank Covariance Matrix Estimation

In this section, we consider a structured model where $\mathbf{\Sigma} = \operatorname{cov}(\boldsymbol{X})$ is approximately low-rank. Using the trace-norm as a convex relaxation of the rank, we propose the following trace-norm penalized optimization program:

$$\widehat{\mathbf{\Sigma}}^{\mathcal{T}}_{2,\gamma} \in \operatorname*{argmin}_{\mathbf{A}\in\mathcal{S}_d}\left\{\frac{1}{2}\|\mathbf{A} - \widehat{\mathbf{\Sigma}}^{\mathcal{T}}_2\|_{\mathrm{F}}^2 + \gamma\|\mathbf{A}\|_{\mathrm{tr}}\right\}, \tag{5.4}$$

where $\mathcal{S}_d$ denotes the set of $d \times d$ positive semi-definite matrices, $\gamma > 0$ is a regularization parameter and $\widehat{\mathbf{\Sigma}}^{\mathcal{T}}_2$, defined in (3.6), serves as a pilot estimator. This trace-penalized method was first proposed by Lounici (2014) with the initial estimator taken to be the sample covariance matrix, and later studied by Minsker (2018) using a different initial estimator. In fact, given the initial estimator $\widehat{\mathbf{\Sigma}}^{\mathcal{T}}_2$, the estimator given in (5.4) has the following closed-form expression (Lounici, 2014):

$$\widehat{\mathbf{\Sigma}}^{\mathcal{T}}_{2,\gamma} = \sum_{k=1}^{d} \max\{\lambda_k(\widehat{\mathbf{\Sigma}}^{\mathcal{T}}_2) - \gamma, 0\} \times \boldsymbol{v}_k(\widehat{\mathbf{\Sigma}}^{\mathcal{T}}_2)\boldsymbol{v}_k(\widehat{\mathbf{\Sigma}}^{\mathcal{T}}_2)^{\mathsf{T}}, \tag{5.5}$$

where $\lambda_1(\widehat{\mathbf{\Sigma}}^{\mathcal{T}}_2) \ge \cdots \ge \lambda_d(\widehat{\mathbf{\Sigma}}^{\mathcal{T}}_2)$ are the eigenvalues of $\widehat{\mathbf{\Sigma}}^{\mathcal{T}}_2$ in an non-increasing order and $\boldsymbol{v}_1(\widehat{\mathbf{\Sigma}}^{\mathcal{T}}_2), \ldots, \boldsymbol{v}_d(\widehat{\mathbf{\Sigma}}^{\mathcal{T}}_2)$ are the associated orthonormal eigenvectors. The following theorem provides a deviation bound for $\widehat{\mathbf{\Sigma}}^{\mathcal{T}}_{2,\gamma}$ under the Frobenius norm. The proof follows directly from Theorem 3.2 and Theorem 1 of Lounici (2014), and therefore is omitted.

THEOREM 5.2. *For any* $t > 0$ *and* $v > 0$ *satisfying* (3.7), *let*

$$\tau = v\sqrt{\frac{m}{\log(2d) + t}} \quad \text{and} \quad \gamma \ge 2v\sqrt{\frac{\log(2d) + t}{m}}.$$

*Then with probability at least* $1 - e^{-t}$, *the trace-penalized estimator* $\widehat{\mathbf{\Sigma}}^{\mathcal{T}}_{2,\gamma}$ *satisfies*

$$\|\widehat{\mathbf{\Sigma}}^{\mathcal{T}}_{2,\gamma} - \mathbf{\Sigma}\|_{\mathrm{F}}^2 \le \inf_{\mathbf{A}\in\mathcal{S}_d}\left[\|\mathbf{\Sigma} - \mathbf{A}\|_{\mathrm{F}}^2 + \min\{4\gamma\|\mathbf{A}\|_{\mathrm{tr}}, 3\gamma^2\operatorname{rank}(\mathbf{A})\}\right]$$

*and*

$$\|\widehat{\mathbf{\Sigma}}^{\mathcal{T}}_{2,\gamma} - \mathbf{\Sigma}\|_2 \le 2\gamma.$$

*In particular, if* $\operatorname{rank}(\mathbf{\Sigma}) \le r_0$, *then with probability at least* $1 - e^{-t}$,

$$\|\widehat{\mathbf{\Sigma}}^{\mathcal{T}}_{2,\gamma} - \mathbf{\Sigma}\|_{\mathrm{F}}^2 \le \min\{4\|\mathbf{\Sigma}\|_2\gamma, 3\gamma^2\}r_0. \tag{5.6}$$

### 5.3 Sparse Precision Matrix Estimation

Our third example is related to sparse precision matrix estimation in high dimensions. Recently, Avella-Medina et al. (2018) showed that minimax optimality is achievable within a larger class of distributions if the sample covariance matrix is replaced by a robust pilot estimator, and also provided a unified theory for covariance and precision matrix estimation based on general pilot estimators. Specifically, Avella-Medina et al. (2018) robustifed the CLIME estimator (Cai, Liu and Luo, 2011) using three different pilot estimators: adaptive Huber, median-of-means and rank-based estimators. Based on the element-wise truncation procedure and the difference of trace (D-trace) loss proposed by Zhang and Zou (2014), we further consider a robust method for estimating the precision matrix $\mathbf{\Theta}^* = \mathbf{\Sigma}^{-1}$ under sparsity, which represents a useful complement to the methods developed by Avella-Medina et al. (2018).

The advantage of using the D-trace loss is that it automatically results a symmetric solution. Specifically, using the element-wise truncated estimator $\widehat{\mathbf{\Sigma}}^{\mathcal{T}}_1 = \widehat{\mathbf{\Sigma}}^{\mathcal{T}}_1(\mathbf{\Gamma})$ in (3.3) as an initial estimate of $\mathbf{\Sigma}$, we propose to solve

$$\widehat{\mathbf{\Theta}} \in \operatorname*{argmin}_{\mathbf{\Theta}\in\mathbb{R}^{d\times d}}\left\{\underbrace{\frac{1}{2}\langle\mathbf{\Theta}^2, \widehat{\mathbf{\Sigma}}^{\mathcal{T}}_1\rangle - \operatorname{tr}(\mathbf{\Theta})}_{\mathcal{L}(\mathbf{\Theta})} + \lambda\|\mathbf{\Theta}\|_{\ell_1}\right\}, \tag{5.7}$$

where $\|\mathbf{\Theta}\|_{\ell_1} = \sum_{k\neq\ell}|\Theta_{k\ell}|$ for $\mathbf{\Theta} = (\Theta_{k\ell})_{1\le k,\ell\le d}$. For simplicity, we write $\mathcal{L}(\mathbf{\Theta}) = \langle\mathbf{\Theta}^2, \widehat{\mathbf{\Sigma}}^{\mathcal{T}}_1\rangle - \operatorname{tr}(\mathbf{\Theta})$.

Zhang and Zou (2014) imposed a positive definiteness constraint on $\mathbf{\Theta}$, and proposed an alternating direction method of multipliers (ADMM) algorithm to solve the constrained D-trace loss minimization. However, with the positive definiteness constraint, the ADMM algorithm at each iteration computes the singular value decomposition of a $d \times d$ matrix, and therefore is computationally intensive for large-scale data. In (5.7), we impose no constraint on $\mathbf{\Theta}$ primarily for computational simplicity.

Before presenting the main theorem, we need to introduce an assumption on the restricted eigenvalue of the Hessian matrix of $\mathcal{L}(\mathbf{\Theta})$. The Hessian can be written as

$$\mathbf{H}_{\mathbf{\Gamma}} = \frac{1}{2}(\mathbf{I} \otimes \widehat{\mathbf{\Sigma}}_1^{\mathcal{T}} + \widehat{\mathbf{\Sigma}}_1^{\mathcal{T}} \otimes \mathbf{I}),$$

where $\mathbf{\Gamma}$ is the tuning parameter matrix in (3.3). For matrices $\mathbf{A}, \mathbf{B} \in \mathbb{R}^{d^2 \times d^2}$, we define $\langle \mathbf{A}, \mathbf{A} \rangle_{\mathbf{B}} = \operatorname{vec}(\mathbf{A})^{\mathsf{T}} \mathbf{B} \operatorname{vec}(\mathbf{A})$, where $\operatorname{vec}(\mathbf{A})$ designates the $d^2$-dimensional vector concatenating the columns of $\mathbf{A}$. Let $\mathcal{S} = \operatorname{supp}(\mathbf{\Theta}^*) \subseteq \{1, \ldots, d\}^2$, the support set of $\mathbf{\Theta}^*$.

DEFINITION 5.1 (Restricted eigenvalue for matrices). For any $\xi > 0$ and $m \geq 1$, we define the maximal and minimal restricted eigenvalues of the Hessian matrix $\mathbf{H}_{\mathbf{\Gamma}}$ as

$$\kappa_-(\mathbf{\Gamma}, \xi, m) = \inf_{\mathbf{W}} \Bigg\{ \frac{\langle \mathbf{W}, \mathbf{W} \rangle_{\mathbf{H}_{\mathbf{\Gamma}}}}{\|\mathbf{W}\|_{\mathrm{F}}^2} : \mathbf{W} \in \mathbb{R}^{d \times d}, \mathbf{W} \neq 0, \exists J \text{ such that } \mathcal{S} \subseteq J, |J| \leq m, \|\mathbf{W}_{J^c}\|_{\ell_1} \leq \xi \|\mathbf{W}_J\|_{\ell_1} \Bigg\};$$

$$\kappa_+(\mathbf{\Gamma}, \xi, m) = \sup_{\mathbf{W}} \Bigg\{ \frac{\langle \mathbf{W}, \mathbf{W} \rangle_{\mathbf{H}_{\mathbf{\Gamma}}}}{\|\mathbf{W}\|_{\mathrm{F}}^2} : \mathbf{W} \in \mathbb{R}^{d \times d}, \mathbf{W} \neq 0, \exists J \text{ such that } \mathcal{S} \subseteq J, |J| \leq m, \|\mathbf{W}_{J^c}\|_{\ell_1} \leq \xi \|\mathbf{W}_J\|_{\ell_1} \Bigg\}.$$

CONDITION 5.1 (Restricted eigenvalue condition). *We say restricted eigenvalue condition with* $(\mathbf{\Gamma}, 3, k)$ *holds if* $0 < \kappa_- = \kappa_-(\mathbf{\Gamma}, 3, k) \leq \kappa_+(\mathbf{\Gamma}, 3, k) = \kappa_+ < \infty$.

Condition 5.1 is a form of the localized restricted eigenvalue condition (Fan et al., 2018). Moreover, we assume that the true precision matrix $\mathbf{\Theta}^*$ lies in the following class of matrices:

$$\mathcal{U}(s, M) = \Bigg\{ \mathbf{\Omega} \in \mathbb{R}^{d \times d} : \mathbf{\Omega} = \mathbf{\Omega}^{\mathsf{T}}, \mathbf{\Omega} \succ 0, \|\mathbf{\Omega}\|_1 \leq M, \sum_{k,\ell} I(\Omega_{k\ell} \neq 0) \leq s \Bigg\}.$$

A similar class of precision matrices has been studied in the literature; see, for example, Zhang and Zou (2014), Cai, Ren and Zhou (2016) and Sun et al. (2018). Recall the definition of $\mathbf{V}$ in Theorem 3.1. We are ready to present the main result, with the proof deferred to the supplementary material.

THEOREM 5.3. *Assume that* $\mathbf{\Theta}^* = \mathbf{\Sigma}^{-1} \in \mathcal{U}(s, M)$. *Let* $\mathbf{\Gamma} \in \mathbb{R}^{d \times d}$ *be as in Theorem 3.1 and let* $\lambda$ *satisfy*

$$\lambda = 4C\|\mathbf{V}\|_{\max} \sqrt{\frac{2\log d + \log \delta^{-1}}{\lfloor n/2 \rfloor}} \quad \text{for some } C \geq M.$$

*Assume Condition 5.1 is fulfilled with* $k = s$ *and* $\mathbf{\Gamma}$ *specified above. Then with probability at least* $1 - 2\delta$, *we have*

$$\|\widehat{\mathbf{\Theta}} - \mathbf{\Theta}^*\|_F \leq 6C\kappa_-^{-1}\|\mathbf{V}\|_{\max} s^{1/2} \sqrt{\frac{2\log d + \log \delta^{-1}}{\lfloor n/2 \rfloor}}.$$

REMARK 5. The nonasymptotic probabilistic bound in Theorem 5.3 is established under the assumption that Condition 5.1 holds. It can be shown that Condition 5.1 is satisfied with high probability as long as the coordinates of $\boldsymbol{X}$ have bounded fourth moments. The proof is based on an argument similar to the proof of Lemma 4 in the work of Sun, Zhou and Fan (2019), and thus is omitted here.

## 6. NUMERICAL STUDY

In this section, we assess the numerical performance of proposed tail-robust covariance estimators. We consider the *element-wise truncated covariance estimator* $\widehat{\mathbf{\Sigma}}_1^{\mathcal{T}}$ defined in (3.3), the *spectrum-wise truncated covariance estimator* $\widehat{\mathbf{\Sigma}}_2^{\mathcal{T}}$ defined in (3.6), the *Huber-type M-estimator* $\widehat{\mathbf{\Sigma}}_1^{\mathcal{H}}$ given in (3.13) and the *adaptive Huber M-estimator* $\widehat{\mathbf{\Sigma}}_3^{\mathcal{H}}$ in Section 4.2.

Throughout this section, we let $\{\tau_{k\ell}\}_{1 \leq k,\ell \leq d} = \tau$ for $\widehat{\mathbf{\Sigma}}_1^{\mathcal{H}}$. To compute $\widehat{\mathbf{\Sigma}}_2^{\mathcal{T}}$ and $\widehat{\mathbf{\Sigma}}_1^{\mathcal{H}}$, the robustification parameter $\tau$ is selected by five-fold cross-validation. The robustification parameters $\{\tau_{k\ell}\}_{1 \leq k,\ell \leq d}$ for $\widehat{\mathbf{\Sigma}}_1^{\mathcal{T}}$ are tuned by solving the equation (4.2), and thus is an adaptive elementwise-truncated estimator. To implement the *adaptive Huber M-estimator* $\widehat{\mathbf{\Sigma}}_3^{\mathcal{H}}$, we calibrate $\{\tau_{k\ell}\}_{1 \leq k,\ell \leq d}$ and estimate $\{\sigma_{k\ell}\}_{1 \leq k,\ell \leq d}$ simultaneously by solving the equation system (4.3) as described in Algorithm 1.

We first generate a data matrix $\mathbf{Y} \in \mathbb{R}^{n\times d}$ with rows being i.i.d. vectors from a distribution with mean $\mathbf{0}$ and covariance matrix $\mathbf{I}_d$. We then rescaled the data and set $\mathbf{X} = \mathbf{Y}\mathbf{\Sigma}^{1/2}$ as the final data matrix, where $\mathbf{\Sigma} \in \mathbb{R}^{d\times d}$ is a structured covariance matrix. We consider four distribution models outlined below:

(1) (Normal model). The rows of $\mathbf{Y}$ are i.i.d. generated from the standard normal distribution.
(2) (Student's $t$ model). $\mathbf{Y} = \mathbf{Z}/\sqrt{3}$, where the entries of $\mathbf{Z}$ are i.i.d. with Student's distribution with 3 degrees of freedom.
(3) (Pareto model). $\mathbf{Y} = 4\mathbf{Z}/3$, where the entries of $\mathbf{Z}$ are i.i.d. with Pareto distribution with shape parameter 3 and scale parameter 1.
(4) (Log-normal model). $\mathbf{Y} = \exp\{0.5 + \mathbf{Z}\}/(e^3 - e^2)$, where the entries of $\mathbf{Z}$ are i.i.d. with standard normal distribution.

The covariance matrix $\mathbf{\Sigma}$ has one of the following three forms:

(a) (Diagonal structure). $\mathbf{\Sigma} = \mathbf{I}_d$;
(b) (Equal correlation structure). $\sigma_{k\ell} = 1$ for $k = \ell$ and $\sigma_{k\ell} = 0.5$ when $k \neq \ell$;
(c) (Power decay structure). $\sigma_{k\ell} = 1$ for $k = \ell$ and $\sigma_{k\ell} = 0.5^{|k-\ell|}$ when $k \neq \ell$.

In each setting, we choose $(n, d)$ as $(50, 100)$, $(50, 200)$ and $(100, 200)$, and simulate 200 replications for each scenario. The performance is assessed by the relative mean error (RME) under spectral, max or Frobenius norm:

$$\mathrm{RME} = \frac{\sum_{i=1}^{200} \|\widehat{\mathbf{\Sigma}}_i - \mathbf{\Sigma}\|_{2,\max,\mathrm{F}}}{\sum_{i=1}^{200} \|\widetilde{\mathbf{\Sigma}}_i - \mathbf{\Sigma}\|_{2,\max,\mathrm{F}}},$$

where $\widehat{\mathbf{\Sigma}}_i$ is the estimate of $\mathbf{\Sigma}$ in the $i$th simulation using one of the four robust methods and $\widetilde{\mathbf{\Sigma}}_i$ denotes the sample covariance estimate that serves as a benchmark. The smaller the RME is, the more improvement the robust method achieves.

Tables 1–3 summarize the simulation results, which indicate that all the robust estimators outperform the sample covariance matrix by a visible margin when data are generated from a heavy-tailed or an asymmetric distribution. On the other hand, the proposed estimators perform almost as well as the sample covariance matrix when the data follows a normal distribution, indicating high efficiencies in this case. The performances of the four robust estimators are comparable in all scenarios: the *spectrum-wise truncated covariance estimator* $\widehat{\mathbf{\Sigma}}_2^{\mathcal{T}}$ has the smallest RME under spectral norm, while the other three estimators perform better under max and Frobenius norms. This outcome is inline with our intuition discussed in Section 3. Furthermore, the computationally efficient *adaptive Huber M-estimator* $\widehat{\mathbf{\Sigma}}_3^{\mathcal{H}}$ performs comparably as the *Huber-type M-estimator* $\widehat{\mathbf{\Sigma}}_1^{\mathcal{H}}$ where the robustification parameters are chosen by cross-validation.

## 7. DISCUSSION

In this paper, we surveyed and unified selected recent results on covariance estimation for heavy-tailed distributions. More specifically, we proposed element-wise and spectrum-wise truncation techniques to robustify the sample covariance matrix. The robustness, referred to as the *tail robustness*, is demonstrated by finite-sample deviation analysis in the presence of heavy-tailed data: the proposed estimators achieve exponential-type deviation bounds under mild moment conditions. We emphasize that the tail robustness is different from the classical notion of robustness that is often characterized by the breakdown point (Hampel, 1971). Nevertheless, it does not provide any information on the convergence properties of an estimator, such as consistency and efficiency. Tail robustness is a concept that combines robustness, consistency, and finite-sample error bounds.

We discussed three types of procedures in Section 3: truncation-based methods, their $M$-estimation counterparts and the median-of-means method. Truncated estimators have closed-form expressions and therefore are easy to implement in practice. The corresponding $M$-estimators achieve comparable sub-Gaussian-type error bounds, which are of the order $\sqrt{\log(d/\delta)/n}$ under the max norm and of order $\sqrt{\mathrm{r}(\mathbf{\Sigma})\log(d/\delta)/n}$ under the spectral norm, but with sharper moment-dependent constants. Computationally, $M$-estimators can be efficiently evaluated via gradient descent method or iteratively reweighted least squares algorithm. Both truncated and $M$-estimators involve robustification parameters that need to be calibrated to fit the noise level of the problem. Adaptation and tuning of these parameters are discussed in Section 4. The MOM estimator proposed in Section 3.4 is tuning-free because the number of blocks depends neither on noise level nor on confidence level. Following the terminology proposed by Devroye et al. (2016), truncation-based estimators are $\delta$-dependent estimators as they depend on the confidence level $1 - \delta$ at which one aims to control, while the MOM estimator achieves sub-Gaussian error bounds simultaneously at all confidence levels in

TABLE 1
*RME under diagonal structure*

| | Normal | | | $t_3$ | | | Pareto | | | Log-normal | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | 2 | max | F | 2 | max | F | 2 | max | F | 2 | max | F |
| | | | | | | $n = 50,\ p = 100$ | | | | | | |
| $\widehat{\mathbf{\Sigma}}_1^{\mathcal{H}}$ | 0.97 | 0.95 | 0.98 | 0.37 | 0.39 | 0.65 | 0.27 | 0.21 | 0.47 | 0.27 | 0.21 | 0.51 |
| $\widehat{\mathbf{\Sigma}}_3^{\mathcal{H}}$ | 0.97 | 0.90 | 0.96 | 0.37 | 0.36 | 0.59 | 0.29 | 0.24 | 0.45 | 0.24 | 0.19 | 0.49 |
| $\widehat{\mathbf{\Sigma}}_1^{\mathcal{T}}$ | 0.97 | 0.91 | 0.96 | 0.40 | 0.38 | 0.62 | 0.27 | 0.23 | 0.42 | 0.25 | 0.18 | 0.50 |
| $\widehat{\mathbf{\Sigma}}_2^{\mathcal{T}}$ | 0.96 | 0.99 | 0.98 | 0.34 | 0.41 | 0.67 | 0.26 | 0.25 | 0.44 | 0.25 | 0.26 | 0.56 |
| | | | | | | $n = 50,\ p = 200$ | | | | | | |
| $\widehat{\mathbf{\Sigma}}_1^{\mathcal{H}}$ | 0.98 | 0.95 | 0.98 | 0.32 | 0.29 | 0.60 | 0.29 | 0.23 | 0.41 | 0.24 | 0.20 | 0.43 |
| $\widehat{\mathbf{\Sigma}}_3^{\mathcal{H}}$ | 0.98 | 0.96 | 0.97 | 0.31 | 0.26 | 0.54 | 0.27 | 0.20 | 0.42 | 0.24 | 0.19 | 0.38 |
| $\widehat{\mathbf{\Sigma}}_1^{\mathcal{T}}$ | 0.97 | 0.95 | 0.96 | 0.33 | 0.29 | 0.63 | 0.26 | 0.19 | 0.39 | 0.23 | 0.18 | 0.42 |
| $\widehat{\mathbf{\Sigma}}_2^{\mathcal{T}}$ | 0.95 | 0.98 | 0.95 | 0.31 | 0.33 | 0.65 | 0.24 | 0.26 | 0.48 | 0.22 | 0.23 | 0.48 |
| | | | | | | $n = 100,\ p = 200$ | | | | | | |
| $\widehat{\mathbf{\Sigma}}_1^{\mathcal{H}}$ | 0.99 | 0.98 | 0.99 | 0.40 | 0.47 | 0.58 | 0.46 | 0.49 | 0.51 | 0.32 | 0.20 | 0.47 |
| $\widehat{\mathbf{\Sigma}}_3^{\mathcal{H}}$ | 0.95 | 0.99 | 0.98 | 0.39 | 0.47 | 0.59 | 0.45 | 0.45 | 0.48 | 0.28 | 0.21 | 0.49 |
| $\widehat{\mathbf{\Sigma}}_1^{\mathcal{T}}$ | 0.97 | 0.94 | 0.97 | 0.38 | 0.45 | 0.57 | 0.46 | 0.49 | 0.51 | 0.26 | 0.27 | 0.47 |
| $\widehat{\mathbf{\Sigma}}_2^{\mathcal{T}}$ | 0.94 | 1.01 | 0.95 | 0.33 | 0.51 | 0.64 | 0.42 | 0.53 | 0.61 | 0.28 | 0.27 | 0.58 |

Mean relative errors of the the four robust estimators $\widehat{\mathbf{\Sigma}}_1^{\mathcal{H}}$, $\widehat{\mathbf{\Sigma}}_3^{\mathcal{H}}$, $\widehat{\mathbf{\Sigma}}_1^{\mathcal{T}}$ and $\widehat{\mathbf{\Sigma}}_2^{\mathcal{T}}$ over 200 replications when the true covariance matrix has a diagonal structure. 2, max and F denote the spectral, max and Frobenius norms, respectively.

TABLE 2
*RME under equal correlation structure*

| | Normal | | | $t_3$ | | | Pareto | | | Log-normal | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | 2 | max | F | 2 | max | F | 2 | max | F | 2 | max | F |
| | | | | | | $n = 50,\ p = 100$ | | | | | | |
| $\widehat{\mathbf{\Sigma}}_1^{\mathcal{H}}$ | 0.97 | 0.94 | 0.97 | 0.68 | 0.12 | 0.68 | 0.68 | 0.23 | 0.59 | 0.58 | 0.27 | 0.46 |
| $\widehat{\mathbf{\Sigma}}_3^{\mathcal{H}}$ | 0.96 | 0.95 | 0.96 | 0.69 | 0.15 | 0.64 | 0.62 | 0.21 | 0.59 | 0.52 | 0.27 | 0.44 |
| $\widehat{\mathbf{\Sigma}}_1^{\mathcal{T}}$ | 0.97 | 0.96 | 0.97 | 0.67 | 0.14 | 0.67 | 0.64 | 0.22 | 0.57 | 0.59 | 0.28 | 0.47 |
| $\widehat{\mathbf{\Sigma}}_2^{\mathcal{T}}$ | 0.95 | 0.99 | 1.02 | 0.56 | 0.26 | 0.71 | 0.62 | 0.27 | 0.60 | 0.50 | 0.33 | 0.51 |
| | | | | | | $n = 50,\ p = 200$ | | | | | | |
| $\widehat{\mathbf{\Sigma}}_1^{\mathcal{H}}$ | 0.97 | 0.94 | 0.98 | 0.77 | 0.21 | 0.76 | 0.67 | 0.34 | 0.50 | 0.69 | 0.23 | 0.67 |
| $\widehat{\mathbf{\Sigma}}_3^{\mathcal{H}}$ | 1.00 | 0.97 | 0.98 | 0.77 | 0.22 | 0.73 | 0.63 | 0.31 | 0.50 | 0.70 | 0.23 | 0.68 |
| $\widehat{\mathbf{\Sigma}}_1^{\mathcal{T}}$ | 0.99 | 0.97 | 0.96 | 0.78 | 0.24 | 0.71 | 0.63 | 0.33 | 0.46 | 0.70 | 0.23 | 0.68 |
| $\widehat{\mathbf{\Sigma}}_2^{\mathcal{T}}$ | 0.95 | 0.98 | 1.00 | 0.74 | 0.35 | 0.80 | 0.61 | 0.34 | 0.51 | 0.66 | 0.31 | 0.72 |
| | | | | | | $n = 100,\ p = 200$ | | | | | | |
| $\widehat{\mathbf{\Sigma}}_1^{\mathcal{H}}$ | 1.00 | 0.96 | 0.99 | 0.79 | 0.23 | 0.78 | 0.63 | 0.46 | 0.57 | 0.53 | 0.21 | 0.47 |
| $\widehat{\mathbf{\Sigma}}_3^{\mathcal{H}}$ | 0.98 | 0.98 | 0.97 | 0.79 | 0.24 | 0.79 | 0.69 | 0.48 | 0.58 | 0.57 | 0.22 | 0.48 |
| $\widehat{\mathbf{\Sigma}}_1^{\mathcal{T}}$ | 1.00 | 1.00 | 0.99 | 0.78 | 0.21 | 0.77 | 0.65 | 0.45 | 0.57 | 0.55 | 0.23 | 0.50 |
| $\widehat{\mathbf{\Sigma}}_2^{\mathcal{T}}$ | 0.97 | 1.02 | 1.03 | 0.73 | 0.32 | 0.83 | 0.62 | 0.54 | 0.61 | 0.50 | 0.29 | 0.55 |

TABLE 3
*RME under power decay structure*

| | Normal | | | $t_3$ | | | Pareto | | | Log-normal | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | 2 | max | F | 2 | max | F | 2 | max | F | 2 | max | F |
| | | | | | | $n = 50, p = 100$ | | | | | | |
| $\widehat{\mathbf{\Sigma}}_1^{\mathcal{H}}$ | 0.98 | 0.95 | 0.98 | 0.58 | 0.30 | 0.71 | 0.48 | 0.29 | 0.57 | 0.69 | 0.39 | 0.79 |
| $\widehat{\mathbf{\Sigma}}_3^{\mathcal{H}}$ | 0.95 | 0.95 | 0.93 | 0.58 | 0.28 | 0.72 | 0.48 | 0.26 | 0.58 | 0.70 | 0.39 | 0.78 |
| $\widehat{\mathbf{\Sigma}}_1^{\mathcal{T}}$ | 0.97 | 0.98 | 0.96 | 0.59 | 0.30 | 0.71 | 0.49 | 0.26 | 0.57 | 0.72 | 0.39 | 0.77 |
| $\widehat{\mathbf{\Sigma}}_2^{\mathcal{T}}$ | 0.98 | 0.98 | 0.99 | 0.52 | 0.33 | 0.77 | 0.47 | 0.31 | 0.60 | 0.66 | 0.45 | 0.81 |
| | | | | | | $n = 50, p = 200$ | | | | | | |
| $\widehat{\mathbf{\Sigma}}_1^{\mathcal{H}}$ | 0.98 | 0.95 | 0.97 | 0.58 | 0.30 | 0.71 | 0.48 | 0.29 | 0.57 | 0.69 | 0.39 | 0.79 |
| $\widehat{\mathbf{\Sigma}}_3^{\mathcal{H}}$ | 0.96 | 0.93 | 0.95 | 0.56 | 0.29 | 0.66 | 0.49 | 0.26 | 0.55 | 0.72 | 0.38 | 0.77 |
| $\widehat{\mathbf{\Sigma}}_1^{\mathcal{T}}$ | 0.98 | 0.97 | 0.97 | 0.59 | 0.27 | 0.71 | 0.48 | 0.26 | 0.58 | 0.70 | 0.36 | 0.80 |
| $\widehat{\mathbf{\Sigma}}_2^{\mathcal{T}}$ | 0.98 | 0.98 | 1.01 | 0.54 | 0.24 | 0.76 | 0.41 | 0.31 | 0.60 | 0.68 | 0.42 | 0.82 |
| | | | | | | $n = 100, p = 200$ | | | | | | |
| $\widehat{\mathbf{\Sigma}}_1^{\mathcal{H}}$ | 0.99 | 0.98 | 1.00 | 0.45 | 0.25 | 0.66 | 0.42 | 0.31 | 0.54 | 0.48 | 0.35 | 0.62 |
| $\widehat{\mathbf{\Sigma}}_3^{\mathcal{H}}$ | 0.98 | 0.98 | 0.99 | 0.47 | 0.26 | 0.68 | 0.41 | 0.30 | 0.53 | 0.47 | 0.34 | 0.61 |
| $\widehat{\mathbf{\Sigma}}_1^{\mathcal{T}}$ | 1.00 | 0.99 | 1.00 | 0.50 | 0.30 | 0.68 | 0.41 | 0.34 | 0.56 | 0.49 | 0.38 | 0.64 |
| $\widehat{\mathbf{\Sigma}}_2^{\mathcal{T}}$ | 0.99 | 1.04 | 1.01 | 0.41 | 0.31 | 0.70 | 0.40 | 0.39 | 0.59 | 0.43 | 0.43 | 0.69 |

a certain range but requires slightly stronger assumptions, namely, the existence of sixth moments instead of fourth.

Three examples discussed in Section 5 illustrate that both element-wise and spectrum-wise truncated covariance estimators can serve as building blocks for a variety of estimation problems in high dimensions. A natural question is whether one can construct a single robust estimator that achieves exponentially fast concentration both element-wise and spectrum-wise, that is, satisfies the results in Theorems 3.1 and 3.2 simultaneously. Here we discuss a theoretical solution to this question. In fact, one can arbitrarily pick one element, denoted as $\widehat{\mathbf{\Sigma}}^{\mathcal{T}}$, from the collection of matrices

$$\mathcal{H} = \Bigg\{ \mathbf{S} \in \mathbb{R}^{d\times d} : \mathbf{S} = \mathbf{S}^{\intercal}, \\ \|\widehat{\mathbf{\Sigma}}_2^{\mathcal{T}} - \mathbf{S}\|_2 \le 2v\sqrt{\frac{\log(2d) + \log\delta^{-1}}{m}} \\ \text{and } \|\widehat{\mathbf{\Sigma}}_1^{\mathcal{T}} - \mathbf{S}\|_{\max} \le 2\|\mathbf{V}\|_{\max}\sqrt{\frac{2\log d + \log\delta^{-1}}{m}} \Bigg\}.$$

Due to Theorems 3.1 and 3.2, with probability at least $1-3\delta$, the set $\mathcal{H}$ is non-empty since it contains the true covariance matrix $\mathbf{\Sigma}$. Therefore, it follows from the the triangle inequality that the inequalities

$$\|\widehat{\mathbf{\Sigma}}^{\mathcal{T}} - \mathbf{\Sigma}\|_2 \le 4v\sqrt{\frac{\log(2d) + \log\delta^{-1}}{m}} \quad \text{and}$$

$$\|\widehat{\mathbf{\Sigma}}^{\mathcal{T}} - \mathbf{\Sigma}\|_{\max} \le 4\|\mathbf{V}\|_{\max}\sqrt{\frac{2\log d + \log\delta^{-1}}{m}}$$

hold simultaneously with probability at least $1-3\delta$.

## ACKNOWLEDGMENTS

The authors would like to thank the referees, Associate Editor and Editor for constructive suggestions that led to an improved paper. S. Minsker is supported by NSF Grant DMS-1712956, Z. Ren is supported by NSF Grant DMS-1812030, Q. Sun is supported by a Connaught Award and NSERC Grant RGPIN-2018-06484 and W.-X. Zhou acknowledges support from NSF Grant DMS-1811376.

## SUPPLEMENTARY MATERIAL

**Supplement to "User-Friendly Covariance Estimation for Heavy-Tailed Distributions"** (DOI: 10.1214/19-STS711SUPP; .pdf). In this supplement, we provide proofs of all the theoretical results in the main text. In addition, we investigate robust covariance estimation and inference under factor models, which might be of independent interest.

## REFERENCES


AVELLA-MEDINA, M., BATTEY, H. S., FAN, J. and LI, Q. (2018). Robust estimation of high-dimensional covariance and precision matrices. *Biometrika* **105** 271–284. MR3804402

BICKEL, P. J. and LEVINA, E. (2008a). Regularized estimation of large covariance matrices. *Ann. Statist.* **36** 199–227. MR2387969

BICKEL, P. J. and LEVINA, E. (2008b). Covariance regularization by thresholding. *Ann. Statist.* **36** 2577–2604. MR2485008

BROWNLEES, C., JOLY, E. and LUGOSI, G. (2015). Empirical risk minimization for heavy-tailed losses. *Ann. Statist.* **43** 2507–2536. MR3405602

BUTLER, R. W., DAVIES, P. L. and JHUN, M. (1993). Asymptotics for the minimum covariance determinant estimator. *Ann. Statist.* **21** 1385–1400. MR1241271

CAI, T., LIU, W. and LUO, X. (2011). A constrained $\ell_1$ minimization approach to sparse precision matrix estimation. *J. Amer. Statist. Assoc.* **106** 594–607. MR2847973

CAI, T. T., REN, Z. and ZHOU, H. H. (2016). Estimating structured high-dimensional covariance and precision matrices: Optimal rates and adaptive estimation. *Electron. J. Stat.* **10** 1–59. MR3466172

CAI, T. T., ZHANG, C.-H. and ZHOU, H. H. (2010). Optimal rates of convergence for covariance matrix estimation. *Ann. Statist.* **38** 2118–2144. MR2676885

CATONI, O. (2012). Challenging the empirical mean and empirical variance: A deviation study. *Ann. Inst. Henri Poincaré Probab. Stat.* **48** 1148–1185. MR3052407

CATONI, O. (2016). PAC-Bayesian bounds for the Gram matrix and least squares regression with a random design. Preprint. Available at arXiv:1603.05229.

CHEN, M., GAO, C. and REN, Z. (2018). Robust covariance and scatter matrix estimation under Huber's contamination model. *Ann. Statist.* **46** 1932–1960. MR3845006

CHEN, X. and ZHOU, W.-X. (2019). Robust inference via multiplier bootstrap. *Ann. Statist.* To appear. Available at arXiv:1903.07208.

CHERAPANAMJERI, Y., FLAMMARION, N. and BARTLETT, P. L. (2019). Fast mean estimation with sub-Gaussian rates. Preprint. Available at arXiv:1902.01998.

CONT, R. (2001). Empirical properties of asset returns: Stylized facts and statistical issues. *Quant. Finance* **1** 223–236.

DAVIES, L. (1992). The asymptotics of Rousseeuw's minimum volume ellipsoid estimator. *Ann. Statist.* **20** 1828–1843. MR1193314

DEVROYE, L., LERASLE, M., LUGOSI, G. and OLIVEIRA, R. I. (2016). Sub-Gaussian mean estimators. *Ann. Statist.* **44** 2695–2725. MR3576558

EKLUND, A., NICHOLS, T. E. and KNUTSSON, H. (2016). Cluster failure: Why fMRI inferences for spatial extent have inflated false-positive rates. *Proc. Natl. Acad. Sci. USA* **113** 7900–7905.

FAN, J., LIAO, Y. and LIU, H. (2016). An overview of the estimation of large covariance and precision matrices. *Econom. J.* **19** C1–C32. MR3501529

FAN, J., LIU, H., SUN, Q. and ZHANG, T. (2018). I-LAMM for sparse learning: Simultaneous control of algorithmic complexity and statistical error. *Ann. Statist.* **46** 814–841. MR3782385

FAN, J., SUN, Q., ZHOU, W.-X. and ZHU, Z. (2019). Principal component analysis for big data. *Wiley StatsRef: Statistics Reference Online*. To appear. DOI:10.1002/9781118445112.stat08122.

HALL, P., KAY, J. W. and TITTERINGTON, D. M. (1990). Asymptotically optimal difference-based estimation of variance in nonparametric regression. *Biometrika* **77** 521–528. MR1087842

HAMPEL, F. R. (1971). A general qualitative definition of robustness. *Ann. Math. Stat.* **42** 1887–1896. MR0301858

HOPKINS, S. B. (2018). Mean estimation with sub-Gaussian rates in polynomial time. Preprint. Available at arXiv:1809.07425.

HSU, D. and SABATO, S. (2016). Loss minimization and parameter estimation with heavy tails. *J. Mach. Learn. Res.* **17** Paper No. 18, 40. MR3491112

HUBER, P. J. (1964). Robust estimation of a location parameter. *Ann. Math. Stat.* **35** 73–101. MR0161415

HUBERT, M., ROUSSEEUW, P. J. and VAN AELST, S. (2008). High-breakdown robust multivariate methods. *Statist. Sci.* **23** 92–119. MR2431867

LEPSKI, O. V. and SPOKOINY, V. G. (1997). Optimal pointwise adaptive methods in nonparametric estimation. *Ann. Statist.* **25** 2512–2546. MR1604408

LEPSKIĬ, O. V. (1990). A problem of adaptive estimation in Gaussian white noise. *Theory Probab. Appl.* **35** 454–466.

LIU, R. Y. (1990). On a notion of data depth based on random simplices. *Ann. Statist.* **18** 405–414. MR1041400

LIU, Y. and REN, Z. (2018). Minimax estimation of large precision matrices with bandable Cholesky factor. Preprint. Available at arXiv:1712.09483.

LIU, L., HAWKINS, D. M., GHOSH, S. and YOUNG, S. S. (2003). Robust singular value decomposition analysis of microarray data. *Proc. Natl. Acad. Sci. USA* **100** 13167–13172. MR2016727

LOUNICI, K. (2014). High-dimensional covariance matrix estimation with missing observations. *Bernoulli* **20** 1029–1058. MR3217437

LUGOSI, G. and MENDELSON, S. (2019). Sub-Gaussian estimators of the mean of a random vector. *Ann. Statist.* **47** 783–794. MR3909950

MARONNA, R. A. (1976). Robust $M$-estimators of multivariate location and scatter. *Ann. Statist.* **4** 51–67. MR0388656

MENDELSON, S. and ZHIVOTOVSKIY, N. (2018). Robust covariance estimation under $L_4$–$L_2$ norm equivalence. Preprint. Available at arXiv:1809.10462.

MINSKER, S. (2015). Geometric median and robust estimation in Banach spaces. *Bernoulli* **21** 2308–2335. MR3378468

MINSKER, S. (2018). Sub-Gaussian estimators of the mean of a random matrix with heavy-tailed entries. *Ann. Statist.* **46** 2871–2903. MR3851758

MINSKER, S. and STRAWN, N. (2017). Distributed statistical estimation and rates of convergence in normal approximation. Preprint. Available at arXiv:1704.02658.

MINSKER, S. and WEI, X. (2018). Robust modifications of $U$-statistics and applications to covariance estimation problems. Preprint. Available at arXiv:1801.05565.

MITRA, R. and ZHANG, C.-H. (2014). Multivariate analysis of nonparametric estimates of large correlation matrices. Preprint. Available at arXiv:1403.6195.

MIZERA, I. (2002). On depth and deep points: A calculus. *Ann. Statist.* **30** 1681–1736. MR1969447

NEMIROVSKY, A. S. and YUDIN, D. B. (1983). *Problem Complexity and Method Efficiency in Optimization. A Wiley-Interscience Publication*. Wiley, New York. MR0702836

PORTNOY, S. and HE, X. (2000). A robust journey in the new millennium. *J. Amer. Statist. Assoc.* **95** 1331–1335. MR1825288

PURDOM, E. and HOLMES, S. P. (2005). Error distribution for gene expression data. *Stat. Appl. Genet. Mol. Biol.* **4** Art. 16. MR2170432

RICE, J. (1984). Bandwidth choice for nonparametric regression. *Ann. Statist.* **12** 1215–1230. MR0760684

ROUSSEEUW, P. J. (1984). Least median of squares regression. *J. Amer. Statist. Assoc.* **79** 871–880. MR0770281

ROUSSEEUW, P. and YOHAI, V. (1984). Robust regression by means of S-estimators. In *Robust and Nonlinear Time Series Analysis* (*Heidelberg*, 1983). *Lect. Notes Stat.* **26** 256–272. Springer, New York. MR0786313

SALIBIAN-BARRERA, M. and ZAMAR, R. H. (2002). Bootstrapping robust estimates of regression. *Ann. Statist.* **30** 556–582. MR1902899

SUN, Q., ZHOU, W.-X. and FAN, J. (2019). Adaptive Huber regression. *J. Amer. Statist. Assoc.* DOI:10.1080/01621459.2018.1543124.

SUN, Q., TAN, K. M., LIU, H. and ZHANG, T. (2018). Graphical nonconvex optimization via an adaptive convex relaxation. In *Proceedings of the* 35*th International Conference on Machine Learning* **80** 4810–4817.

TUKEY, J. W. (1975). Mathematics and the picturing of data. In *Proceedings of the International Congress of Mathematicians, Vancouver, 1975*, vol. 2.

TYLER, D. E. (1987). A distribution-free $M$-estimator of multivariate scatter. *Ann. Statist.* **15** 234–251. MR0885734

VERSHYNIN, R. (2012). Introduction to the non-asymptotic analysis of random matrices. In *Compressed Sensing* 210–268. Cambridge Univ. Press, Cambridge. MR2963170

WANG, L., ZHENG, C., ZHOU, W. and ZHOU, W.-X. (2018). A new principle for tuning-free Huber regression. Technical Report.

YOHAI, V. J. (1987). High breakdown-point and high efficiency robust estimates for regression. *Ann. Statist.* **15** 642–656. MR0888431

ZHANG, T., CHENG, X. and SINGER, A. (2016). Marčenko–Pastur law for Tyler's $M$-estimator. *J. Multivariate Anal.* **149** 114–123. MR3507318

ZHANG, T. and ZOU, H. (2014). Sparse precision matrix estimation via lasso penalized D-trace loss. *Biometrika* **101** 103–120. MR3180660

ZUO, Y. and SERFLING, R. (2000). General notions of statistical depth function. *Ann. Statist.* **28** 461–482. MR1790005